The Development of Branch Site Selection Models

Capstone 2 - Final Project Report
Florida Polytechnic University
Spring 2024

submitted by:
MidFlorida Branch Site Selection
Madison Yonash – Business Analytics
Claire Robinson – Data Science
Anthony Nguyen – Business Analytics
Emily Crowley – Applied Mathematics
Andrei Bailes – Business Analytics
Justine Saylor – Computer Science
Susan Nunez – Business Analytics
Alexia Belden – Business Analytics

submitted to:
Dr. James Dewey
May 1 2024

1. Introduction

This project is sponsored by MidFlorida Credit Union with the goal of determining sites where a new branch is most likely to succeed. MidFlorida is a credit union that was founded in Polk County and has spread to have branches as northern as Gainesville and as southern as Naples and Stewart. To achieve the project goal, the team used a mixture of different approaches to produce viable sites for MidFlorida. The determination of potentially profitable sites is structured using game theory to tie the different components together. This process begins with the identification of census tracts that have potential profitability for MidFlorida. This is done by cross-referencing a market saturation analysis that determines what census tracts in Florida are underserved and a Heckman model that estimates the number of checking accounts for MidFlorida in each census tract. Within these census tracts, heat maps pull together a variety of different relevant factors to determine specific areas that have potential profitability. These factors include spatial competitive analysis, web scraping analysis for other businesses indicating profitability, and traffic factors. Ultimately, the team was able to produce three proof of concept locations for MidFlorida.

Table 1. Appendix numbers and relevant content

Appendix	Content
1	Full list of customer needs, requirements and design objectives
2	Concepts and Analysis of Alternatives
3	Game theory
4	Heckman selection
5	Market saturation analysis
6	Web scraping analysis
7	Heat maps and site recommendations

2. Customer Needs, Objectives and Team Interpretation

To ensure prompt completion of relevant tasks and to break down the project into more manageable components, the team broke into sub teams to tackle the project. This overarching task of determining what makes a potential branch site successful, and what locations have those features, was divided into multiple objectives for teams of 1-2 members to work on. To achieve this, the team met regularly as a whole and weekly with their mentor.

2.1 Customer Needs

MidFlorida needs a better way to model profitability and select potential sites for new branch locations in Florida using publicly available and in-house data. The result must be meaningful in context of MidFlorida's requirements and have high interpretability so that upper management can understand it. As a final deliverable, the team submitted a comprehensive folder and potential sites. See Appendix 1 for a list of primary, secondary, and latent customer needs.

2.2 Design Objectives

The main goal is to give MidFlorida a clear framework for choosing new sites and utilize these determinants to identify potential new site locations. The design criteria evaluate the customer's subjective perceptions of what makes a site attractive to the decision-makers at MidFlorida. For example, when thinking of criteria for a good site, someone may say that the location must be in a "convenient location." This can be quantified using web scraping to determine what nearby

businesses MidFlorida members may travel to and utilize or by adding a component that considers travel time for MidFlorida members. The design should include subcomponents that are complimentary; the web scraping results should build on the information provided in the traffic analysis. The heat maps must be able to visually tie together the different components of the modeling process.

2.3 Team Interpretation

Discrepancy existed between the sponsor's and the faculty mentor's interpretations of what needed to be done to complete the project. The sponsor suggested making improvements to the pre-existing model, while the faculty mentor argued that the model was fundamentally flawed because of its reliance on current branch sites. The description of the original model is found in Appendix 1, following the customer needs. The sub teams created to address the customer needs include teams working on a Heckman model, market saturation, web scraping, spatial analysis, and visualizations. These analyses are performed concurrently to draw conclusions about potential site viability.

2.4 Functional and Non-functional Requirements

The main functional requirement of this project is to provide suggestions for new branch locations. This meant creating models that estimate profitability for MidFlorida and determining what factors of a location are indicators of a profitable site. This requires working models that produce useful results in selection without using biased data in the determination of these results. Each modeling component for determining branch viability (market saturation analysis, Heckman modeling, etc.) is a functional requirement. Other requirements included data cleaning, data visualization, and model evaluation. Non-functional requirements include the associated reports and documentation of the processes and models created. The use of game theory may be considered a non-functional component, in that it provides a framework for the analysis to be performed but does not provide insight on its own. Other non-functional requirements include interpretability and interactivity of heat maps.

2.5 Limitations

MidFlorida's reliance on their current model for determining potential profitability (Appendix 1) created a need to completely recreate the modeling process. A limitation was that MidFlorida was unable to share certain data due to privacy issues, such as addresses of customers. Many of the more precise and useful web scraping tools were blocked by a paywall, resulting in the need to utilize free tools instead. The computational requirements to create heat maps containing large amounts of data and to run computationally expensive models were difficult on the devices owned by team members. Additionally, learning new techniques while simultaneously implementing them provided challenges in understanding the "why" behind what was being done.

3. Concept Generation and Analysis of Alternatives

Literature review was performed to understand what alternatives are available to MidFlorida and to determine what work has been previously done in this field. These ideas were discussed with MidFlorida, and the limitations of existing alternatives were explained to the team. The concept that was ultimately developed resulted in the reworking of initial ideas that developed with the progression of the project. Game theory provided a framework in which the other components of modeling fit. Different concepts and models addressed different necessary components of determining profitability: both for MidFlorida and for general financial institutions.

3.1 Literature Search

Literature review was performed extensively throughout the development of the team's analytical approaches. Different companies offer services for site selection for other businesses. These services

can be costly and produce unhelpful results (for example, suggesting a location in the middle of the woods). See Appendix 2 for list of sources and the literature search results.

3.2 Analytical Approach Generation

Due to the selection bias, MidFlorida's current method of prediction was not statistically sound and new methods had to be generated. The approaches generated follow the framework of game theory as described in Section 4.1.1 and Appendix 3. First, the market saturation model (Section 4.1.3 and Appendix 5) predicts the probability of a census tract being underserved and the Heckman selection model (Section 4.1.2 and Appendix 4) predicts the number of checking accounts as a measure of profitability. Spatial analysis (Section 4.1.5) is performed, and web scraping (Section 4.1.4 and Appendix 6) examines what businesses near a branch make it a convenient location for customers. The mapping team (Section 4.1.6) then uses the leaflet package in R to create interactive heatmaps that tie all the solution paths and analysis together in an easy to understand visual.

3.3 Analysis of Alternatives

Due to the complicated statistical nature of this problem, no single approach will work on its own because of the compounding and intertwined bias issues of the available data. A list of the analysis of alternatives, considering what each component adds to determination of profitability and the limitations they each have, is included in Appendix 2.

It was decided that the current best approach would be to use the different models in a complimentary fashion with each one assisting in covering the flaws of the others. As the problem continues to be explored and other potential solutions are determined, the criteria for evaluating them revolves around one main question: Does this new approach improve or add to the current system? The only way to answer this question is to try the new approach out and then compare its results to the other models. If the new path can give additional insight into potential locations, then it can either be added to the current system or replace one of the current models (see Appendix 2 for Table 1: Evaluation Criteria and Figure 1: Diagram of How Current Evaluation Criteria is Applied).

3.4 Flow Chart of Design Process

The team decided to utilize different models concurrently to determine where profitable potential branch locations may be located. From a high level, the design process that has been utilized is shown in Figure 1 of Appendix 2. As potential solutions paths were determined, the only way to rule them out was to test them and compare how they did to the other solution paths.

4. Prototyping Process

As previously described, the developed method for determining potentially profitable branch site locations for MidFlorida has many moving parts. Since the work of the different sub-teams is interconnected, an effort was made to start the prototyping process as soon as possible. As these separate approaches were refined, the separate indicators for potential profitability were updated. This allowed for a better understanding of the methods used and gave insights for further investigation.

4.1 Proof of Concepts

The framework defining the approach for determining a potentially profitable location for MidFlorida detailed in this report is through game theory. Discussed in section 4.1.1, structuring the different approaches through a game theory framework brings the separate ideas together harmoniously into a final deliverable. Section 4.1.6 ties together the different components into a heat map that functions as a visual tool to better understand the final deliverable. Section 4.1.2. includes

further reasoning behind why MidFlorida's current approach is not statistically sound using a Heckman selection model. The active approach created by this team begins with the market saturation model, detailed in 4.1.3. After which methods of web scraping and spatial analysis described in 4.1.4 and 4.1.5. were implemented for continued investigation.

4.1.1 Proof of Concepts – Game Theory

Game theory was utilized as the overarching structure for analysis of the project. This structure considers the different census tracts in Florida and the different financial institutions present, with the strategy of each financial institution being where they had branch locations. Within this structure, the team aims to estimate both the number of financial institutions in each census tract at market equilibrium (to determine underserved areas) and MidFlorida's strategy (where they should put branches to maximize profit). This is discussed thoroughly in Appendix 3.

4.1.2 Proof of Concepts – Heckman Selection

In a perfect world, sites would be completely randomly selected, allowed to run for ten years, and their profitability and demographics would be compared. Unfortunately, as discussed in Section 4.1.1, businesses have a strategy and will place branches where they think they can maximize their profits. When sites are not selected that had good observable values, on average they had bad unobservables. When sites are selected that had bad observable values, on average they had good unobservables. Due to this selection bias, any model that does not account for it will underpredict the effects of the observables on the outcome because the selection forces a negative relationship between the observables and unobservables. With this in mind, the team made Heckman selection models that combine a selection modeling step and an outcome modeling step that takes selection bias into account when making predictions. There is no way to know for sure if the Heckman model was able to fully remove the selection bias present in the data as there is no "true" value to test against for tracts where there is no branch. However, the Heckman model was able to show and prove the role that selection plays in the data as shown in Figure 2 of Appendix 4. Appendix 4 contains a comprehensive report on the Heckman theory and results as well as explaining the proof of concept in further detail.

4.1.3 Proof of Concepts – Market Saturation

As mentioned in 4.1.1, the first step in the team's process for creating a recommendation for MidFlorida is to create a model that will estimate the number of financial institutions at market equilibrium for each census tract in Florida, the sum of values for each column in matrix B*. This market saturation analysis is performed assuming that the market is at equilibrium and that free entry exists for the financial market (see Appendix 3, free entry). A Poisson regression model was created because it is best for count data, which is what is used here, and it prevents the predictions from being negative. The predictor variables for the regression model are demographic information including population density, square miles, bachelor's degree population ratio, elderly population ratio, renter occupied housing ratio, median household income, and owner-occupied with a mortgage housing ratio. From this, the team formulated an indicator for potential profitability using the Poisson distribution. For each census tract, the probability that the market equilibrium number of financial institutions is greater than what currently exists. The census tracts with the highest probabilities were then investigated by the other sub-teams (described in 4.1.4 and 4.1.5) to find specific locations within those census tracts that may be suitable for MidFlorida. Further explanations of the development and design of the market saturation model are in Appendix 5.

4.1.4 Proof of Concepts – Web Scraping

For MidFlorida's specific strategy the team decided to determine a new branch location based on whether the location was convenient for customers. To do this, the team used an automated and manual approach to discover the common franchises around different banks including MidFlorida, Suncoast, Truist, and Wells Fargo. The automated method used the G Maps Extractor web scraping extension to search nearby a branch location using the search categories "restaurants" and "stores". The manual method involved the website Google Map Developers which created a two-mile circle around a branch location the team used to manually create a list of all the businesses within the radius. Based on the results of both methods, the team used a web scraping browser extension Webscraper.io to get all the locations of the selected businesses in Florida. Further explanations of web scraping and the automated/manual methods can be found in Appendix 6.

4.1.5 Proof of Concepts – Spatial Analysis

Not one map can determine the best location for a new site for MidFlorida. The team needs to look at multiple maps. These include population density and member counts, Heckman model heat map, market saturation heat map, web scraping data, and financial institutions. While looking at these maps, it will be best to use all maps that allow us to know the area better. There is no automated way to compare competitive threats and opportunities, so a manual interpretation of spatial opportunities and threats is necessary.

4.1.6 Proof of Concepts – Heat Maps

The team was able to create multiple interactive heat maps when determining a profitable site location for MidFlorida. The first heat map that the team created was the Poisson probability and predicted checking account heat map. The second heat map that the team was able to create was the population density heat map. The third heat map created by the team was the MidFlorida members heat map. The fourth heat map that the team created was the median income heat map. The fifth heat map that the team created was the credit unions heat map. The final heat map created was the banks heat map.

4.2 Design Details

The team created an interactive branch site selection model in RStudio, which contained all the heat maps, daily traffic flow, and color-coded locations and sites. The probability and predicted checking account heat map contain the market saturation model, which shows the probability of a census tract being underserved, and the Heckman model, which shows the predicted checking account for each census tract. The population density and median income heat maps show the population density and median income for each census tract. The MidFlorida members heat map shows how many members of MidFlorida reside in each census tract. The credit unions and banks heat map show the numbers of credit unions and banks in each census tract. The annual average daily traffic flow was color coded to be red with the darker shade of red representing high traffic flow. MidFlorida's locations were color coded to be green, while their credit union competitors were color coded to be red and their bank competitors were color coded to be orange. The Publix's locations obtained by the web scrapping team were color coded to be blue, while other businesses were color coded to be yellow. The branch site recommendations to MidFlorida were color coded to be cyan. Appendix 7 shows examples of the heat maps and their use for determining potentially profitable locations.

4.3 Beta Prototype

The beta prototype for the MidFlorida model includes the market saturation Poisson regression model, the web scraping sitemaps, the Heckman Correction model, and the traffic/spatial analysis which are all combined in the interactive heat maps. From this beta prototype, the team presented three potential locations to MidFlorida in Hillsborough, Pinellas, and Orange County. This prototype does meet customer needs by identifying variables that affect profitability, creating models that can

be cross-referenced to identify potential sites, finding data that is useful for profitability, as well as being simple enough to explain to higher management. From the alpha prototype the team made progress in finishing all the models and combining them on the heat maps to pick a potential site. The only remaining issue is not having a true "test" set to see how well the selected locations perform. The major change from the alpha prototype was the change of the market saturation model. The first version of the regression model was linear. This was changed to a Poisson regression model, which is more suitable for countable data and prevents negative predictions.

5. Results

5.1 Final Product Testing

The team first used the branch site selection model to determine what census tracts were underserved and what its predicted checking accounts were. After identifying census tracts that had a high probability of being underserved and a high predicted checking accounts, the team then performed further analysis by looking at the heat maps within these specific census tracts. The team also examined the competitor locations, web scraping results, and traffic flow of the census tracts. Overall, the branch site selection model met all the design objectives and specifications and was very successful as we were able to find and recommend three branch site locations for MidFlorida. The team received positive feedback from MidFlorida on the usefulness of these recommendations. As aforementioned, the only way to truly "test" the results of these models would be to open a branch site at those locations, but our team does not have the ability to do this.

6. Conclusions and Future Work

Overall, the team was able to achieve the goals that were set at the beginning of the project and provide MidFlorida with a total of three site recommendations in an easy-to-understand visualization. All the sub teams work is comprised of many moving parts that were able to be comprehensively tied together into a final visualization, deliverable, and steps for future work and improvements laid out.

6.1 Conclusions and Project Summary

All the sub teams contributed and put effort into their parts. Game theory was used to help analyze the problem better and figure out how to quantify the strategies of any financial institution of interest, using branch locations. The Heckman model wasn't fully able to remove the selection bias since the development of instrumental variables that are uncorrelated with profitability was difficult to perform. However, it did prove the point of selection playing a role in the model and justify the necessity of using alternative approaches. A Poisson regression model was utilized to determine the market saturation at different census tracts. The census tracts with the highest probabilities of being underserved were given to the other sub teams to further investigate if these areas contain a promising location for MidFlorida. The web scraping team was able to read through the census tracks manually and automatically to obtain the common businesses near successful and non-successful MidFlorida. This allowed the team to help compare growth and decline in MidFlorida branches to find if these businesses were impactful to the growth of MidFlorida or its decline. From this data many businesses were found to have a positive impact on MidFlorida like Publix, Dollar General, Advance Auto Parts, Beef-O-Brady's, and Cheddar's Scratch Kitchen. These businesses were later added to the heat map for better analysis. The spatial analysis sub team needed to analyze the map to find the census tracts had the highest probabilities and checking accounts biased on the Heckman and market saturation model. This would allow the sub team to know the area better by seeing the population, annual average daily traffic, the businesses nearby. It also was a good tool to visual areas that MidFlorida is vulnerable to a competitor. The interactive heat map had multiple layers that pulled everything together and became a key tool in examining promising census tracts. It included population density and members count, probability and predicting checking accounts, traffic, and

web scraping data. The team was able to give MidFlorida three different locations and they were pleased with the results. The team was able to explain the overall process in a concise and understandable manner to the MidFlorida team, both those who were familiar with the project and those who had not previously seen the work of the team.

7.2 Future Work

For future work, the team would recommend adding a temporal component that would greatly improve the overall predictive power of the model. This can be done by determining how the current locations have changed over time and what changes in predictors have caused changes in profitability. The team also recommends identifying better instrumental variables for the Heckman Correction and predicting the market saturation in the next three years. Since the population in Florida is constantly growing along with other changes in demographics, if MidFlorida can predict these changes they can beat competition in expanding areas.

7.3 Reflection Report

If the team was to start this over from scratch, the first thing the team would do differently is change how we divided up the teams. The original teams were census data (Andrei and Kenny), market saturation (Emily and Anthony), transportation analysis (Lexi and Susan), examining current model (Claire and Madi), and web scraping (Justine). By the end of the project the teams were market saturation (Emily), transportation/spatial analysis (Lexi and Susan), heat maps and data collection (Andrei and Anthony), Heckman selection model (Claire), game theory (Madi), and web scraping (Justine). This final division of teams worked very well, as everyone had a more specific part to focus on resulting in three potential locations and multiple models. The heat maps and game theory became critical parts in presenting our models and were not a part of the original teams.

Another thing the team would do differently is to start creating the models sooner. The team was very intimidated at taking the first steps as many members had to start from scratch in learning the different parts of the project. By implementing the models earlier, the team would have had more time to fine-tune and correct the models instead of trying to make them perfect from the start. What worked well was having a starting point, even if the starting point was very far from the final model. After the team started developing the models, they made significant progress from the fall to spring semester. By taking longer on the initial steps the team had to work till the last minute to start fitting all of the different pieces together. The team also would have met more with our mentor at the beginning of the fall semester to better understand the project as he tried to express taking the first steps before, but the team didn't quite understand.

In terms of data, the team would have decided to work with census tracts instead of zip codes from the start. Since the team initially started working with zip codes there were many issues with using the data forcing the change to census tracts. By making this change later, many sub teams had to go back and switch all the previous data, therefore costing more time. For the Heckman selection sub team, they would have started with two-step models instead of maximum likelihood Heckman models. The first model caused a lot of problems with the maximum models not being stable. For the web scraping sub team, they would have started testing the web scraping tools sooner. Many of the preliminary tools that were marketed as "free" turned out to be free only up to a certain point. By finding out the limitations of the different tools sooner, the team could have taken better advantage of the free features that reset every month.

Overall, what worked well was how effectively and efficiently every member of the team worked together. Without every member pulling their weight the team would not have been able to make as much progress as we did in the fall and spring semesters. Having a Discord channel with all members

helped ensure constant communication where everyone was up to date on the project's status. Also having a OneDrive with all the files for the team members to access at any time allowed team members to reference past presentations and reports quickly. With Andrei as the team's point of contact with the sponsor, it made communication with the sponsor easy and created a good relationship with the sponsor. With Madi as the team's leader she was able to keep all the team members on the same page and help lead presentations as well as answer any questions. All the reports and assignments that needed to be turned in were collaborative documents that each team member could work on at the same time. The team met every week on Monday, Wednesday, and Friday in a reserved conference room, therefore everyone was dedicating time each week to working on the project together to help each other and be involved with the final product.

7.4 Broader Impacts

The goal of this project is to find new branch sites for MidFlorida to expand to. As part of this expansion, property could be bought up when it could have been used for something else. For example, where our models show it would be best to open a new branch could also be the best place to open a new community center, park, or food bank. That said, this project could also positively impact the community. The market saturation model predicts census tracts where there should be more financial institutions than there currently are. In other words, it finds underserved areas. Because of that, this project could benefit both MidFlorida and the community by helping to decrease the number of financial deserts in Florida. However, if the model were to work perfectly and cause marginalized areas or vulnerable locations to be underserved, then it could affect accessibility to financial resources and cause the development of financial institution deserts.

8.0 Individual Contributions

The team completed their individual contributions required as needed. All members contributed to the final project.

Team Member	Contribution
Madison Yonash	12.5%
Claire Robinson	12.5%
Anthony Nguyen	12.5%
Emily Crowley	12.5%
Justine Saylor	12.5%
Susan Nunez	12.5%
Alexia Belden	12.5%
Andrei Bailes	12.5%

Appendix 1 - Customer Needs, Requirements and Objectives

1.1 Primary, secondary, latent needs

Primary

- Create models to identify potential branch sites that does so in a way that it can be explained to higher management.
- Identification of variables that affect profitability and the success of a branch.

Secondary

- Scalability to different geographical areas of consideration.
- Data quality checks, including sample selection bias.

Latent

- Balancing a model that can be as accurate as possible while not overfitting.

1.2 Initial Approach to Branch Site Selection by MidFlorida

MidFlorida currently uses data from their current branches to determine if a potential site may be worth opening a new branch. The branch selection process currently begins with someone asking the sponsor to look at a potential location using his model. He will then look either at that area or find intersections and potential sites in this area and focus on those. To utilize the current model, the sponsor must input the different features of the site into an Excel spreadsheet, which creates bar graphs of the different features to compare the potential site to the pre-existing branch locations. These features include things such as population, number of credit unions within a predetermined distance, and median income. For each of the features, the current model allows the sponsor to compare the value of that feature against the high performing branches, fast growing branches, and closed branches. This analysis is performed utilizing a bar graph, and not scored quantitatively. This approach can only examine the potential of one site at a time, and does not allow for easy comparison of different potential sites.

This approach cannot produce statistically sound results, which is discussed more in depth in the Heckman Theory. In the beginning of the determination of the team's analytical approach, the idea was to add new predictors to MidFlorida's current model to improve it. Upon discussion of the model and the issues that make it inaccurate, it was determined that this approach was inappropriate for producing useful outcomes. This is due to the relatively small sample size of locations, and the sample selection bias that exists for these locations, meaning that the data produced from the sample cannot be applied generally outside of the sample.

Appendix 2 - Concepts and Analysis of Alternatives

2.1 Literature Review

Targomo is a company that offers business expansion and site selection help for other companies. They offer a platform called TargomoLOOP that allows the client to use interactive maps, comprehensive data as well as analytics to help with decision-making searching to a new level. Targomo considers transportation with statistical data. The program also considers competitor locations where the client can visually see. Along with the option to upload data this program might have potential utilization for MidFlorida (Business Expansion & Site Selection).

MossAdams offers a service for site selection and gives a general overview of how site selection improves decision-making and provides a list of benefits from a site selection analysis. The website provides an overview of the site selection process and contact information for a primary contact for this offer. Though this service is not a solution because this service will be less effective on a big-scale company (Site selection studies to grow and expand your business).

MapBusinessOnline is a business mapping software that offers a variety of features that can help organize the site selection process and narrow down the options so that a decision-making team can sort through impacting factors. Depending on the extent of the site selection requirement, the client may require several types of maps that work together to narrow down the options. This software might be a beneficial resource (Ives Site selection mapping projects - decision making tools for expansion planning).

Faster Capital has a comprehensive market analysis report that suggests Market saturation is a crucial factor to consider when conducting a comprehensive market analysis. A deeper into the definition and importance of market saturation, providing examples, tips, and case studies to illustrate its significance in a comprehensive market analysis report. Analyzing market saturation is a critical step in conducting a market saturation analysis. (Analyzing market saturation for a comprehensive market analysis report).

The US Department of Transportation offers an extensive amount of data and resources that could aid in research collecting but, in this case, could not be a final product. The D.O.T. Allows for use of their product so it could be applied to other models that allow uploaded data (*Transportation Analysis Tools*).

2.2 Literature Review Sources

"Business Expansion & Site Selection." *Targomo*, 20 Mar. 2023, www.targomo.com/use-cases-overview/business-expansion-site-selection.

"Site Selection Studies to Grow and Expand Your Business." *Site Selection Studies to Grow and Expand Your Business*, www.mossadams.com/services/accounting/tax/credits-and-incentives/site-selection-services. Accessed 27 Apr. 2024.

Ives, Geoffrey. "Site Selection Mapping Projects - Decision Making Tools for Expansion Planning." *MapBusinessOnline Blog*, 17 Aug. 2022, www.mapbusinessonline.com/blog/2022/08/17/site-selection-mapping-projects-decision-making-tools-for-expansion-planning/.

"Analyzing Market Saturation for a Comprehensive Market Analysis Report." *FasterCapital*, fastercapital.com/content/Analyzing-Market-Saturation-for-a-Comprehensive-Market-Analysis-Report.html. Accessed 27 Apr. 2024.

Transportation Analysis Tools. U.S. Department of Transportation. (n.d.). https://www.transportation.gov/grants/dot-navigator/transportation-analysis-tools

2.3 Advantages and Limitations of Modeling Techniques

1. The inherent bias of MidFlorida's current model and its data means that it cannot give truly accurate predictions and leaves out many possibly important factors such as traffic flow and customer commutes around the location. Even using Heckman selection modeling, without having instrumental variables that fully capture selection (without also being a determinant of profitability), this prediction of number of checking accounts is not completely unbiased. The Heckman selection model also is at the granularity of census tracts, meaning it cannot indicate where specifically within a census tract a potential branch site will be successful.
2. A market saturation model helps to understand the ratio of banks/credit unions to people that can support a new branch and help with determining areas to investigate. It determines saturation of a census tract but does not say specifically where in that census tract a branch site would be successful. It is not specific to MidFlorida; therefore, it cannot predict how well a MidFlorida branch will do as it doesn't consider important variables such as traffic flow and factors specific to MidFlorida.
3. Web scraping provides useful insights into what other businesses indicate profitability of financial institutions and profitability specific to MidFlorida. This information provides useful insight in looking at the viability of potential sites and intersections but does not provide information for determining profitability as a quantitative prediction.
4. Customer mapping analysis could provide valuable insight into where the current customers drive, their commutes, and areas where customers are coming from far enough away that MidFlorida risks losing them to other credit unions. While this analysis helps in picking a potential site for consideration and aids in getting branch success predictors, it alone cannot fully determine how well a branch will do. This approach does not consider the predictors and demographic factors of the locations that determine profitability. This approach also is limited for the team as MidFlorida does not have the information about where their current customers work, and further exploration would need to be done to determine the commute flows of these members.
5. Spatial awareness analysis could provide valuable insight into where competitors are located along with businesses that are frequently located close to financial institutions. This analysis could consider where communities are located and any land that would not be favorable. This analysis alone cannot account for all the factors needed to determine how well a site will do.
6. Heat mapping analysis is a viable tool for pulling together the aforementioned analytical tools but cannot function without their existence as purely looking at demographic factors without the consideration of market saturation or predictive models means that the usefulness of the tool is severely limited.

Table 1: Evaluation Criteria

Criteria	Importance	Model Adjustment	Market Saturation Model	Customer Mapping Model	Spatial Awareness	Potential Solution
Value/Insight added	100%	5	5	5	5	?
Overall Score: (Out of 5)	100%	5	5	5	5	?

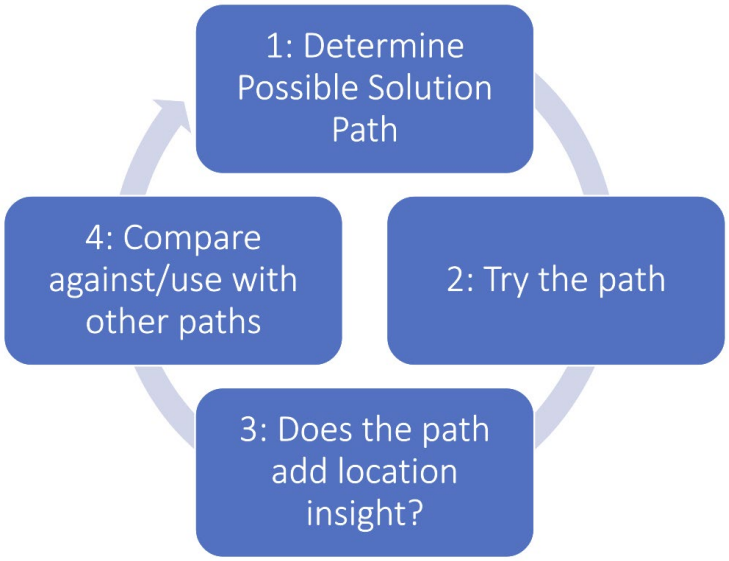
1: Determine Possible Solution Path
2: Try the path
3: Does the path add location insight?
4: Compare against/use with other paths

Figure 1: Diagram of How Current Evaluation Criteria is Applied

Appendix 3 – Game Theory

Abstract
MidFlorida Credit Union is seeking to determine potential sites within Florida that could serve as potentially profitable new branch locations. This paper attempts to serve as a means of explaining the theory behind branch site selection in our capstone project.

1 Game Theory

1.1 Game Theory Overview

In game theory, players make strategic decisions to maximize their benefit within a given situation. It's assumed that all players have access to all available information and will base their decisions on their best estimate of what others will do. Each player's decision is made under the assumption that all others are simultaneously optimizing their own strategies. For example, in terms of profit, denoted as π, with have players A and B, the profit for each player can be described as:

$$\pi_{\mathrm{A}}\left(X_A, X_B\right)$$

$$\pi_{\mathrm{B}}\left(X_B, X_A\right)$$

A Nash Equilibrium is reached when all decisions are optimized simultaneously, considering the actions of others, and no player has an incentive to change their strategy. This is equilibrium is expressed as:

$$X_A^* \rightarrow argmax\ \pi_A\left(X_A, X_B^*\right)$$

$$X_B^* \rightarrow argmax\ \pi_B\left(X_B, X_A^*\right)$$

1.2 Game Theory for MidFlorida

The problem that MidFlorida is trying to solve can be described using game theory. In this example, we can describe the different locations as:

$$g = \{1, 2, \ldots, G\}$$

In the case of our work, each location g can be thought of as a census tract of interest in Florida. Similarly, we can describe the different financial institutions as:

$$i = \{1, 2, \ldots, N\}$$

We can let $i_{\mathrm{MidFlorida}}$ represent our financial institution of interest. The strategy of each financial institution can be defined as the presence or absence of a branch at a certain census tract. For example, for each institution, this could be written in the format:

$$b_i = (0, 1, 1, 1, \ldots, 0)$$

where

$$b_{ig} \begin{cases} 1 & \text{if a branch for institution } i \text{ is open at location } g \\ 0 & \text{else} \end{cases}$$

And the number of components within the strategy is equal to the number of locations, G.

Using this, we can create a matrix that represents all financial institutions and their strategies, which can be defined as B:

$$\begin{bmatrix} \cdots & b_{1g} & \cdots \\ \cdots & b_{2g} & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & b_{Ng} & \cdots \end{bmatrix}$$

Each row represents the strategy of a particular financial institution i, which is their presence at different locations g.
The number of locations with branches for an institution i is the sum of each row of the matrix, defined:

$$n_i = \sum_{g=1}^{G} b_{ig} = \text{Number of branches for financial institution } i$$

See Section 2 for more information on estimating this.

Similarly, the number of institutions with branches in census tract g is the sum of each column of the matrix, defined as:

$$n_g = \sum_{i=1}^{N} b_{ig} = \text{Number of institutions present in census tract } g$$

See section 3 for more information on estimating this.

Considering that the goal for each financial institution is to choose a strategy that optimizes their profit (let π represent profit) a financial institution of interest i must consider the strategies of all other financial institutions. The profit also must consider X, the set of predictors that are considered for the set of locations. X can be described using a matrix of all census tracts g and their relevant features (such as median income, population density, etc). The profit π for a financial institution i can therefore be notated as:

$$\pi_i(b_i, b_{-i}; X)$$

Assuming that matrix B is at a Nash Equilibrium, we can denote it as B*. B* assumes that our environment has the current optimized strategy of each player considering all other players, and that the addition of one more branch location by any institution i at any location g would result in a loss of profit.

1.3 Free Entry
Let the following represent the profit

$$\pi_{jg}(\tilde{b}_{jg}, b_{jg}^{ne})$$

Where π_{jg} is the highest possible profit possible if one more branch opens at location g, and j is the financial institution that would make the most profit at location g.
If the market is at a perfect equilibrium, B*, then adding a branch at any location g would result in a negative profit (loss). The extent of the number of competitors, N, is determined by the nature of entry of the market. The more barriers to entry, the fewer competitors that there will be. With free entry, potential profit is exhausted at an equilibrium, and this permits the estimation of the market saturation analysis (how many financial institutions can each census tract g support). Similarly, certain factors influence the number of branches that can be supported in a location g. This can be written as:

$$\sum_i b_{ig} - B_g(X_g, X_{\sim g})$$

In this B_g is the number of branches open at location g, and B_g (X_g , X_{ng}) is a row of matrix X. For the purposes of this project, we are ignoring $X_{\sim g}$. In matrix X, it is a combination of locations g as the columns and predictors x as the rows.

$$\begin{bmatrix} \cdots & locations & \cdots \\ predictors & \ddots & \vdots \\ \cdots & \cdots & \cdots \end{bmatrix}$$

It is assumed that $X_{\sim g}$ can be achieved as a first approximation of B_g (X_g).

2 Estimating Number of Branches for MidFlorida
For the strategy of MidFlorida, the strategy for the row $i_{MidFlorida}$ is examined. The number of branches that MidFlorida has is:

$$n_i = \sum_{g=1}^{G} b_{(MidFlorida)g} = \text{Number of MidFlorida branches}$$

To determine the strategy of MidFlorida and estimate if the credit union will have a branch at a census tract, the team is exploring different approaches to approximate the strategy of MidFlorida. This involves using a Heckman Correction Model to determine the potential profitability of an area for MidFlorida. Additionally, web scraping is being performed to see what other institutions are present around MidFlorida. This serves as an indicator that a location g that does not have a branch but has those locations can be a potentially profitable branch location. Using transportation analysis can also be an indicator of a successful branch if there are no financial institutions near a high-traffic area.

3 Estimating Number of Financial Institutions per Census Tract
To determine the number of financial institutions per census tract, a market saturation model is being

constructed. The model uses inputs such as number of credit unions, elderly population ratio, median home value, and population density. A key assumption being made to estimate the number of financial intuitions in each census tract is that the markets are being pushed towards equilibrium. This model is essentially estimating each column of the matrix B*, defined as:

$$\sum_{i=1}^{N} b_{ig} = \text{Number of institutions present in census tract } g$$

A high negative residual indicates that the model predicts that a location should have more institutions than it does indicates that the area may be under served.

4 Further Work and Considerations

It is worth nothing that this is created considering the strategy at time t. Over time, the strategy and number of branches for the different financial institutions changes. Adding a temporal component to this analysis would result in a more complete and accurate prediction of potential profitability. For the game theory itself, the framework is general enough to be used for the estimation of any financial institution and in any area.

Appendix 4 - Heckman Selection Model

Abstract:
This is part of the work of the 2023-2024 MidFlorida Branch Site Selection Capstone team at Florida Polytechnic University. This report specifically discusses the work done by the Heckman Selection model sub-team. Heckman models use two separate models– selection and outcome– to try and account for selection bias which is present in MidFlorida's observations. The Heckman models were used to predict the number of potential checking accounts in each census tract in Florida and, in turn, helped automate some of the search process for new sites.

1 Heckman Overview

The Heckman correction was invented by James Heckman in 1974 for the purpose of correcting for selection bias caused by non-randomly selected samples. A two-step Heckman model is broken down into two stages:

1. Stage 1 finds the probability of data entering the sample typically through the use of a Probit regression model.
2. Stage 2 regresses the outcome variable often using ordinary least squares (OLS) regression taking into account the outcome of the selection equation through the inverse Mill's ratio (often referred to as Heckman's lambda).

1.1 What is a Heckman Model?

Generally speaking, some number of predictors *x* causes some dependent variable *y* based on three conditions:

1. The dependent variable follows the independent variable.
2. The dependent changes with the independent variable
3. There is no unobserved factor that would disrupt or get rid of the relationship between the independent variable and the dependent variable.

Data that breaks the 3rd condition can do so for several reasons, one being selection bias within the data itself. Heckman models specifically concentrate on dealing with this type of data.
Consider some unobserved variables Y_{1i}^* and Y_{2i}^* that follow these equations:

$$Y_{1i}^* = X_i\beta + u_i \quad (1)$$
$$Y_{2i}^* = Z_i\lambda + v_i \quad (2)$$

Here, X and Z_i represent sets of independent variables, β and λ are coefficients, and u_i and v_i are some error terms.
Selection occurs for the observed variables $A_{1i}\ and\ A_{2i}$ such that:

$$Observed: Y_{1i} = Y_{1i}^*\ and\ Y_{2i} = 1\ IF\ Y_{2i}^* > threshold \quad (3)$$
$$Unobserved: Y_{1i}\ not\ observed\ and\ Y_{2i} = 0\ IF\ Y_{2i}^* \leq threshold \quad (4)$$

In other words, if the unobserved variable Y_{2i}^* does not surpass some threshold, then that observation will not be selected into the sample ($Y_{2i} = 1$) such that the value for Y_{1i}^* will be observed in Y_{1i}. If you try to regress Y_{1i} on X_i, it will not take into account the effect of selection that is held in the error term v_i meaning that is it not simple random noise.

Heckman Selection models are specifically designed to account for this issue by modeling the selection process as well as its impact on the outcome. They account for observations being chosen for the sample based on unobserved variables Y_{2i}^* and correct the outcome regression for this selection bias.

Selection:

$$Z_i^* = \gamma X_i + \delta W_i + u_i \quad (5)$$

$$Z_i = \begin{cases} 1 \; if \; Z_i^* > 0 \\ 0 \; if \; Z_i^* \leq 0 \end{cases} \quad (6)$$

Where:
Z_i^* *is an unobserved variable the determines propensity for observation*
Z_i *is the binary variable that represents selection* (*Accepted tracts*)
γ *and* δ *are estimated parameter vectors*
X_i *is a vector of explanatory variables*
W_i *is a vector of insrtamental variables*
u_i *is the error term*

Outcome:

$$Y_i = X_i\beta + \delta\lambda\left(\frac{PDF}{CDF}\right) + v_i \quad (7)$$

Where:
Y_i *is the predicted dependent variable* (# *of checking accounts*)
X_i *is a vector of explanatory variables*
β *is an estimated parameter vector*
δ *is the coefficient of the inverse Mill's Ratio*
$\lambda\left(\frac{PDF}{CDF}\right)$ *is the inverse Mill's Ratio* (*a function of the cumulative density function and the probability density function*)
v_i *is the error term*

The diagram below shows a visual representation of how the Heckman model works in the case of MidFlorida's branch site selection problem. The two stages of the Heckman Selection Model are shown: the selection and outcome stages (referring to equations 5-6 and 7, respectively). In the correlations between the predictors and the outcomes of the models, the instrumental variables are represented by bolded arrows. These are predictors that only affect selection and do not affect the outcome.

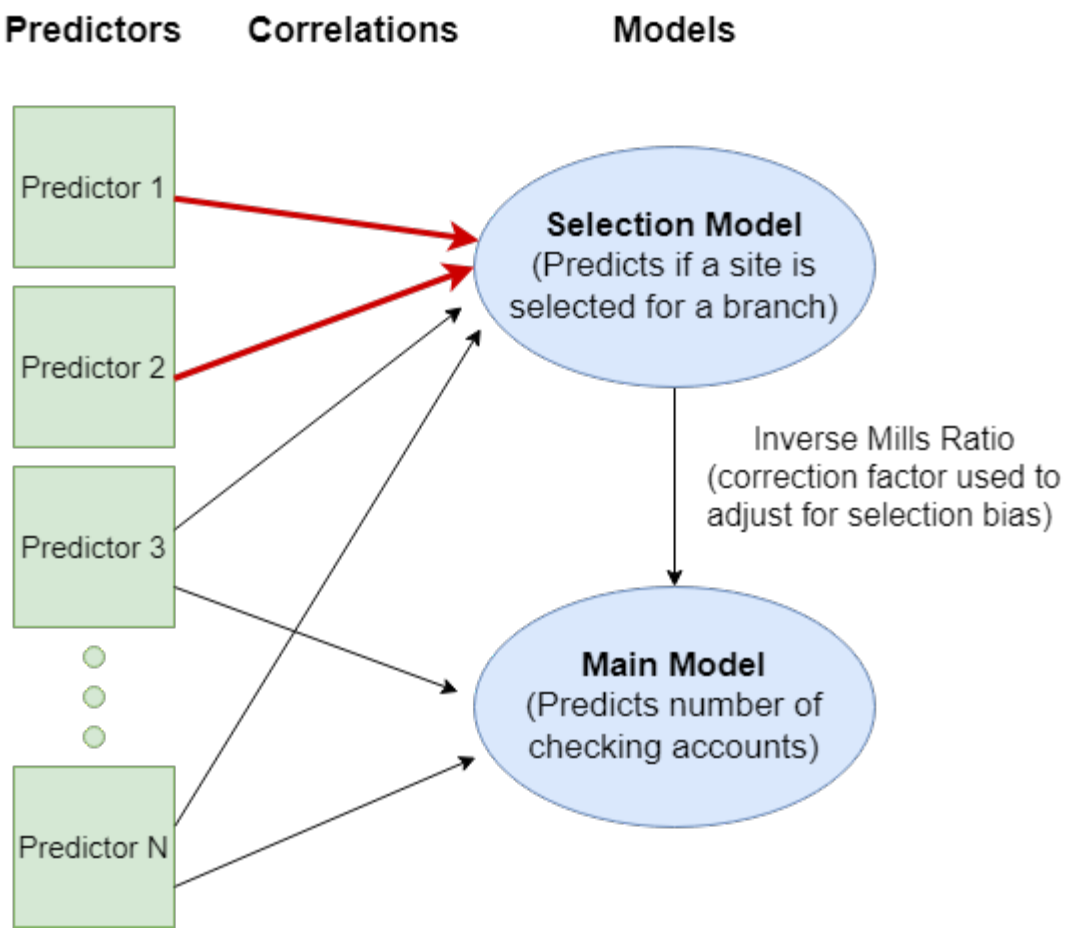

Figure 1: Diagram of Heckman model structure

2 Why Does This Project Need a Heckman Model?

In a perfect world, 100 completely random sites would be selected for MidFlorida branches, allowed to run for about 10 years before their profitability and area demographics are compared. Unfortunately, a company with the goal of making money will not randomly select sites but will instead choose sites where they believe they can maximize their profits– therein lies the problem of selection bias.

In MidFlorida's case, when sites are not selected that had good observable values, on average they had bad unobservables. When sites are selected that had bad observable values, on average they had good unobservables. This type of bias is described in equations 1-4. MidFlorida's current model does not account for the selection bias within their data and, therefore, underpredicts the effects of the observables on the outcome because the selection forces a negative relationship between the observables and unobservables.

3 Results

During the duration of this project, many different models were tested, five of which are listed below. The two tables include the parameter estimates along with the relative standard errors. These Heckman models follow a "twostep" estimation process as opposed to the base maximum likelihood (ML) estimation. The main difference between the two methods is that ML estimation estimates the selection and outcome models simultaneously using maximum likelihood while twostep estimates

the two models separately. Twostep models were used for this project because they tend to be more stable than the ML models.

Selection Model:

Table 1: Parameter estimates and standard errors for five selection models

	Model 1	Model 2	Model 3	Model 4	Model 5
lnPopulation_Density	-0.240	-0.116		-0.053	-0.134
	(0.124)	(0.123)		(0.123)	(0.202)
lnMedian_Home_Value		-0.881	-0.899	-0.553	-0.468
		(0.286)	(0.279)	(0.221)	(0.329)
lnRenter_Ratio	-0.206		-0.200	-0.214	-0.355
	(0.179)		(0.215)	(0.174)	(0.173)
lnfair_share_pop	-0.214	-0.165	-0.229	-0.150	0.265
	(0.174)	(0.147)	(0.160)	(0.164)	(1.471)
lnemployment_ratio	-0.837	-0.690	-0.704		-1.486
	(0.394)	(0.352)	(0.359)		(1.459)
cu_ratio					-2.196
					(5.664)
lnBachelor_Ratio					-0.273
					(0.288)
lnMedian_Household_Income		0.817	0.562		
		(0.399)	(0.488)		
Intercept	12.644	13.237	15.288	16.189	17.373
	(1.708)	(3.142)	(3.959)	(2.615)	(5.879)

Outcome Model:

Table 2: Parameter estimates and standard errors for five outcome models

	Model 1	Model 2	Model 3	Model 4	Model 5
lnPopulation_Density	0.027	0.027	0.042	0.027	0.023
	(0.053)	(0.053)	(0.049)	(0.053)	(0.051)
lnMedian_Home_Value	0.010	0.010		0.010	0.037
	(0.152)	(0.152)		(0.152)	(0.107)
lnfair_share_pop	-0.495	-0.495	-0.447	-0.495	-0.490
	(0.091)	(0.091)	(0.085)	(0.091)	(0.090)
lnemployment_ratio	0.499	0.499	0.369	0.499	0.493
	(0.202)	(0.202)	(0.177)	(0.202)	(0.201)
cu_ratio	1.956	1.956	1.882	1.956	1.957
	(0.171)	(0.171)	(0.161)	(0.171)	(0.171)
lnMedian_Household_Income	0.057	0.057	0.103	0.057	
	(0.218)	(0.218)	(0.148)	(0.218)	
Partial_5mi	-0.173	-0.173	-0.160	-0.173	-0.175
	(0.243)	(0.243)	(0.240)	(0.243)	(0.243)
Intercept	-0.623	-0.623	-1.335	-0.623	-0.331
	(1.866)	(1.866)	(1.789)	(1.866)	(1.527)
lambda	-0.513	-0.321	-0.244	-0.135	-1.535
	(0.310)	(0.299)	(0.304)	(0.302)	(3.456)

Table 3: AIC and BIC for the five potential models

	Model 1	Model 2	Model 3	Model 4	Model 5
AIC	698.53	695.19	701.59	697.44	**694.82**
BIC	796.14	**792.79**	799.5	795.04	805.47

****NOTE:** The AIC and BIC were estimated using maximum likelihood (ML) Heckman models and NOT the twostep models used through the rest of the project because twostep models do not allow estimation of AIC and BIC**

Model 2 has the lowest BIC and model 5 has the lowest AIC. For this, we will use model 5 to explain the results as it has the best AIC and makes the most intuitive sense to explain. The selection model had five predictors and one instrumental variable partial_5mi as shown in equation 8. The subsequent outcome model also has five predictors as well as the value for the inverse Mill's Ratio generated by the selection model.

Selection Model 5:

$$Z_i = \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5 + \delta W_1 + u_i \quad (8)$$

$$\gamma_1 = 0.023 \quad X_1 = Pop\ Density$$
$$\gamma_2 = 0.37 \quad X_2 = Median\ Home\ Value$$
$$\gamma_3 = -0.49 \quad X_3 = Fair\ Share\ Pop$$
$$\gamma_4 = 0.493 \quad X_4 = Employment\ Ratio$$
$$\gamma_5 = 1.957 \quad X_5 = Credit\ Union\ Ratio$$
$$\boldsymbol{W_1 = -0.175 \quad X_2 = Partial\ 5mi}$$

Outcome Model 5:

$$Y_i = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + X_5\beta_5 + X_5\beta_5 + \delta\lambda\left(\frac{PDF}{CDF}\right) + v_i \quad (9)$$

$$\beta_1 = -0.134 \quad X_1 = Pop\ Density$$
$$\beta_2 = -0.468 \quad X_2 = Median\ Home\ Value$$
$$\beta_3 = 0.265 \quad X_3 = Fair\ Share\ Pop$$
$$\beta_4 = -1.486 \quad X_4 = Employment\ Ratio$$
$$\beta_5 = -0.355 \quad X_5 = Renter\ Ratio$$
$$\delta = -1.535 \quad \lambda\left(\frac{PDF}{CDF}\right) = Inverse\ Mill's\ Ratio$$

After choosing a model, we can use it to make predictions for the potential number of checking accounts for each census tract. The graph below shows the distributions for predictions made for census tracts where MidFlorida branches are present– the white bars– and those where they are not– the gray bars.

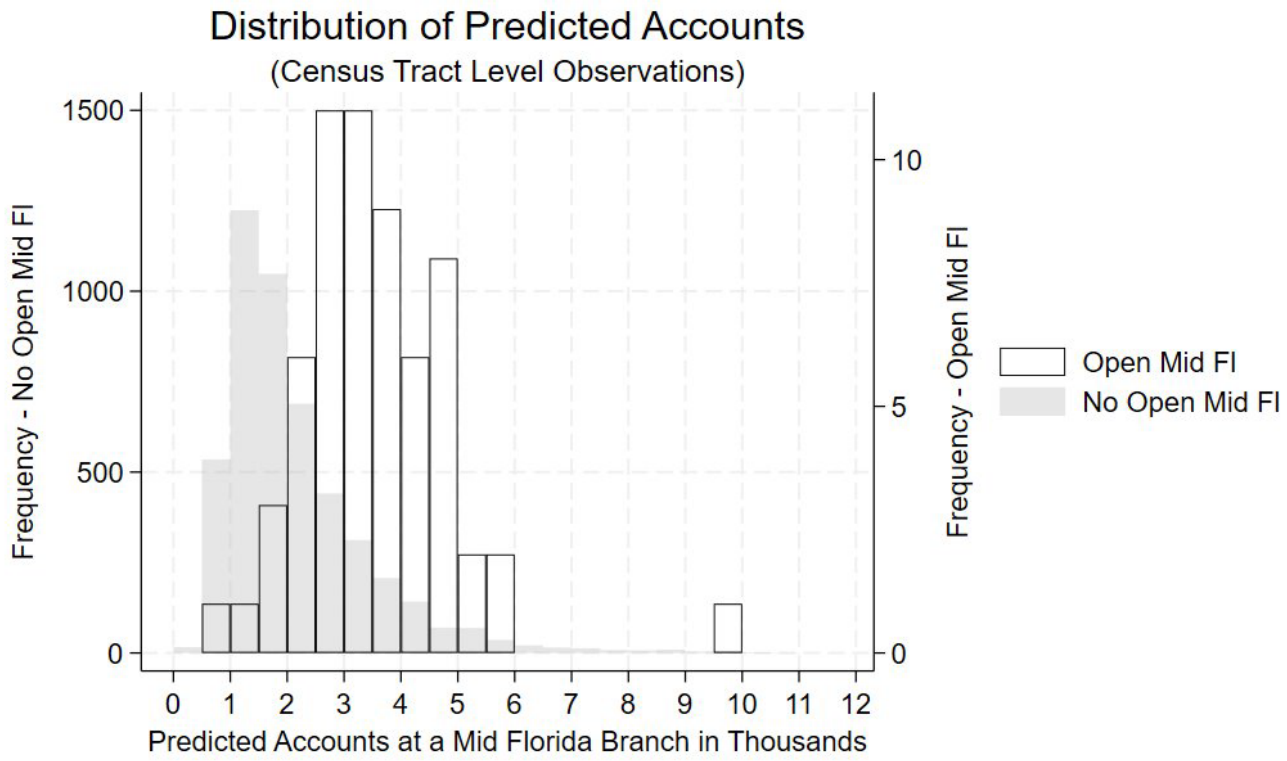

Figure 2: Heckman model predicted checking accounts by census tract (gray = No MidFlorida, white = MidFlorida)

There is a major difference between the distribution of predictions made in the census tracts where MidFlorida branches are present versus those where they are not. Tracts where there are no MidFloridas have systematically lower predictions. This shows that the process of selection plays a big role in how well a branch will do and that selection bias is present in the data.

The predictions generated with the Heckman model can be combined with the work from the other sub-teams such as Web Scraping and Traffic Flow Analysis to create easy-to-understand visualizations in the form of heatmaps such as the one below showing Polk County as an example.

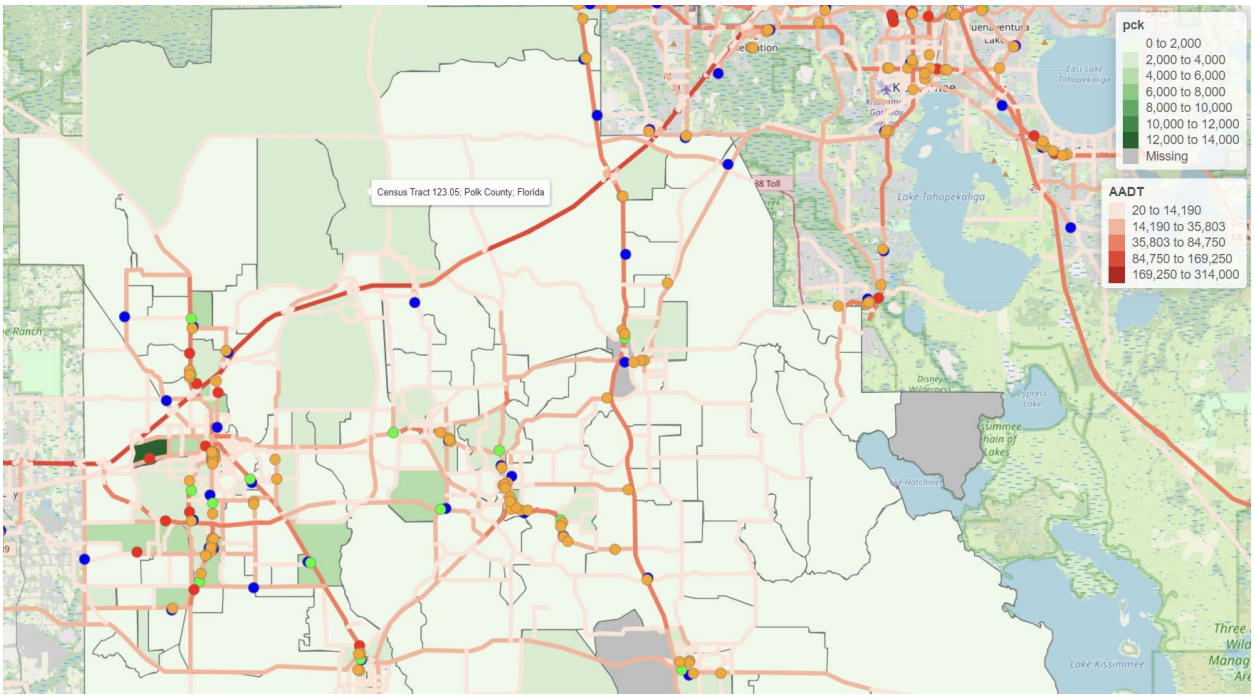

Figure 3: Heckman model predictions tied together with traffic analysis and web scraping in a comprehensive heatmap

It can also be combined with the output from the Market Saturation model to assist in automating the search process. Census tracks with a high predicted number of checking accounts and a high predicted probability of being under-served are shown in dark purple.

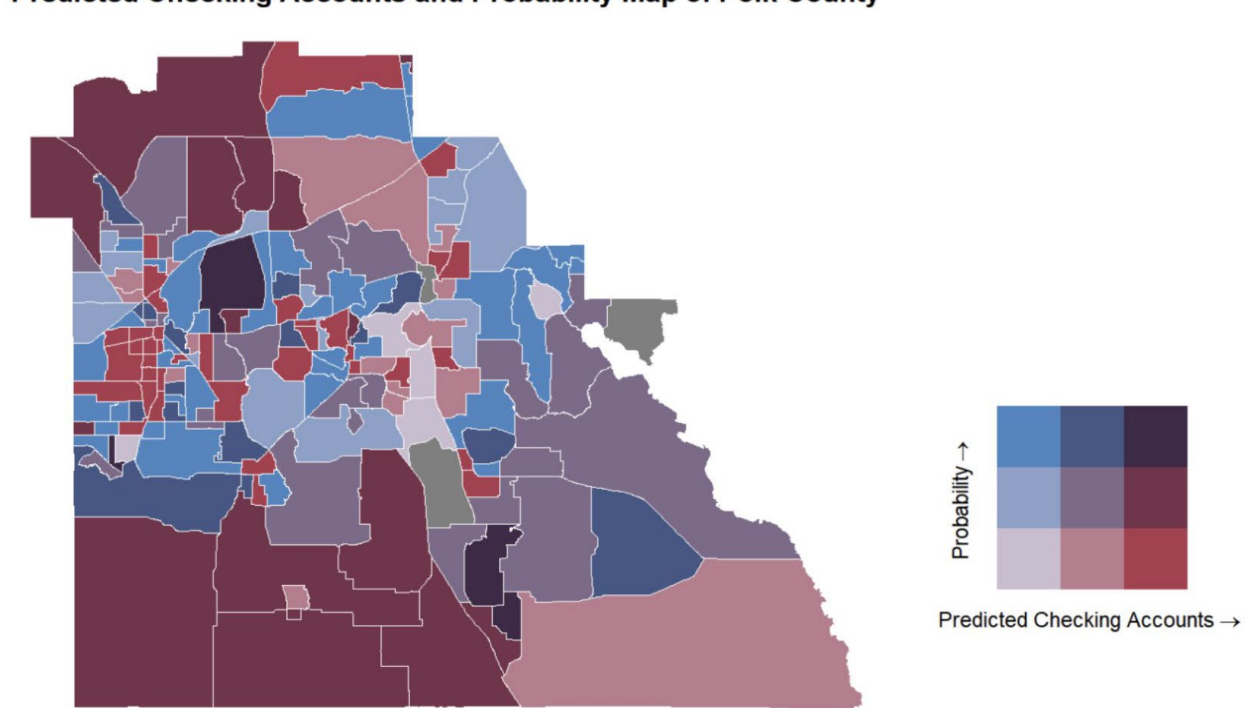

Figure 4: Combined heatmap of Heckman model and Market Saturation model outputs

This narrows down the number of potential census tracts which can then be manually searched using the interactive map created by the Mapping team which combines the work of all the sub-teams into one visualization along with other predictors such as population density.

3.1 Future Works

The Heckman model created can still be improved upon in the future by either MidFlorida or a subsequent Capstone team. The two most pressing improvements are discussed in the following two sections.

3.1.1 Temporal Component

Adding a temporal component would potentially increase the accuracy and predictive power of the Heckman model. For example, a branch that has a large number of checking accounts may have a high number not because it's in the best location but because the branch has been open for a long time. Allowing a temporal component would allow the model to take this into consideration. This has been experimented with in some of the models that were run in the form of either a total count of months or a Boolean of whether or not a branch had been open for more than 5 years but it still needs further adjusting.

Along with this, demographic data and checking account data for every branch would ideally be included in the model for every year from the time they opened. This would allow the team to study how the changing demographics affected the number of checking accounts at each branch. However, this data may not be readily available.

3.1.2 Instrumental Variables

Instrumental variables are imperative for a Heckman model to properly account for selection bias. Specifically, endogeneity is when the outcome variable correlates with the error term of the regression function which causes biased parameter estimates. As instrumental variables only correlate with the predictors and not the error term, Heckman models can use them to effectively separate the influence of selection on the relationship between the predictors and the outcome variable. Wolfolds and Siegal– professors at the University of Michigan– discuss the importance of using instrumental variables in Heckman models that satisfy the condition of only correlating to selection in their paper *Misaccounting for Endogeneity: The Peril of Relying on the Heckman Two-Step Method without a Valid Instrument*.

MidFlorida provided a list of "exclusionary" criteria that they use as a loose guideline when looking at potential sites. These criteria are listed below:

1. A new branch can be no more than 40 miles/1 hour from an existing branch.
2. A new branch cannot be within 5 miles of an existing branch.
3. Have a population of 40,000 within 3 miles of the potential site.
4. Have a population of 3,500 per financial institution within a 3-mile radius of a potential site.

Implementation was attempted for criteria 3 and 4 but as this project worked with census tracks, these criteria didn't translate well. Converting the values to population density also did not work as census tracts vary wildly in size and may show a low population density even with a city in it simply due to the tract size.

For criteria 1 and 2, they were represented as Boolean variables full_5mi, partial_5mi, none_5mi, full_40mi, partial_40mi, and none_40mi that indicated to what degree census tracts were within a 5 or 40 mile radius of existing MidFlorida branches. MidFlorida themselves admit that they do not follow these criteria exactly and many of their current branches are specifically well within 5 miles of one another, especially in heavily populated areas. Because of this, these variables do not have the best correlation to the tracts that were selected. The mapping of these criteria is shown below.

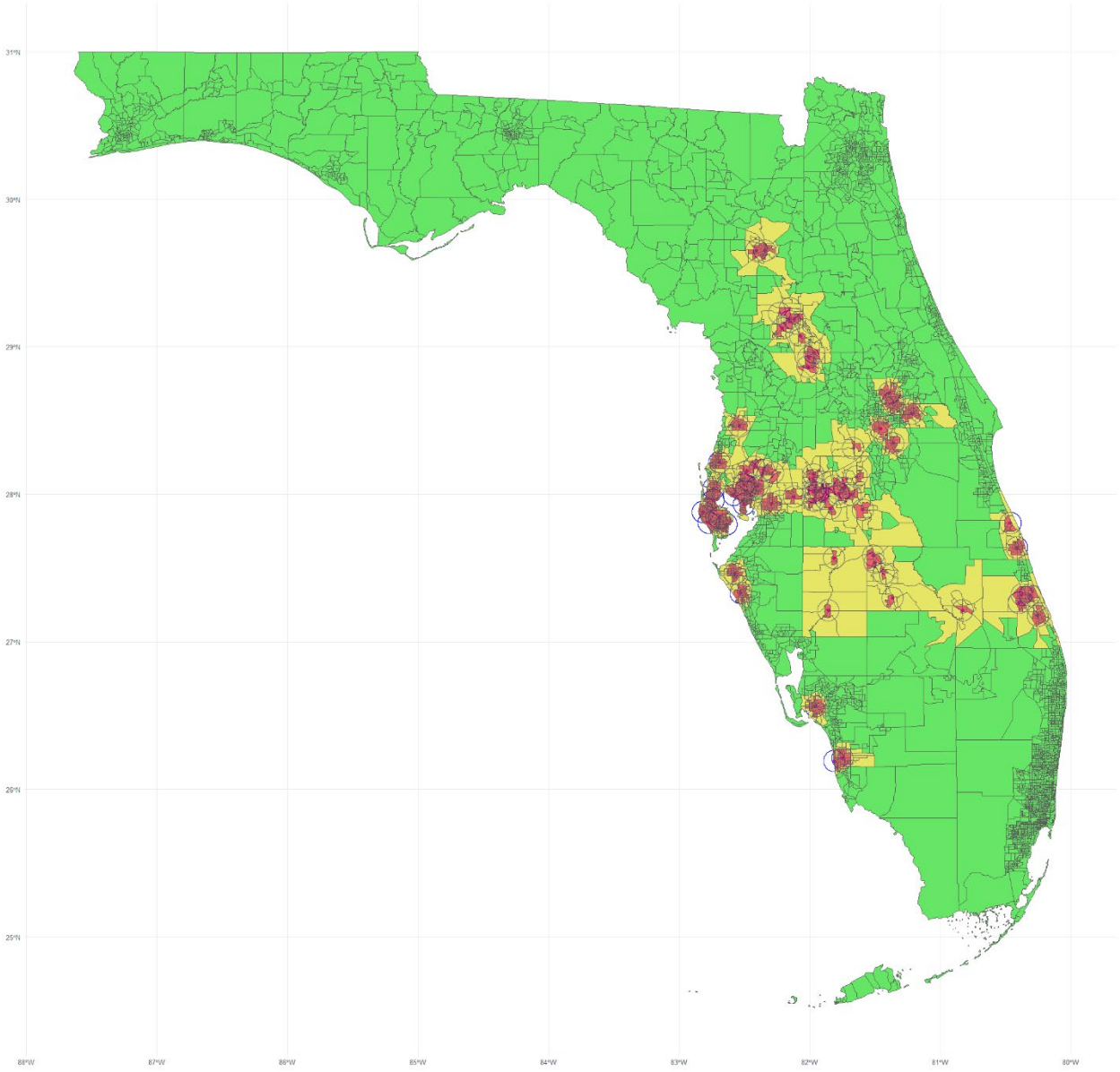

Figure 2: Map of Florida with 5 mile radius circles drawn around existing MidFloridas

Tracts that are not within 5 miles of an existing branch are shown in green, tracts that are partially within are shown in yellow, and fully within are shown in red. In heavily populated areas such as Tampa have circles that overlap with multiple MidFlorida branches.

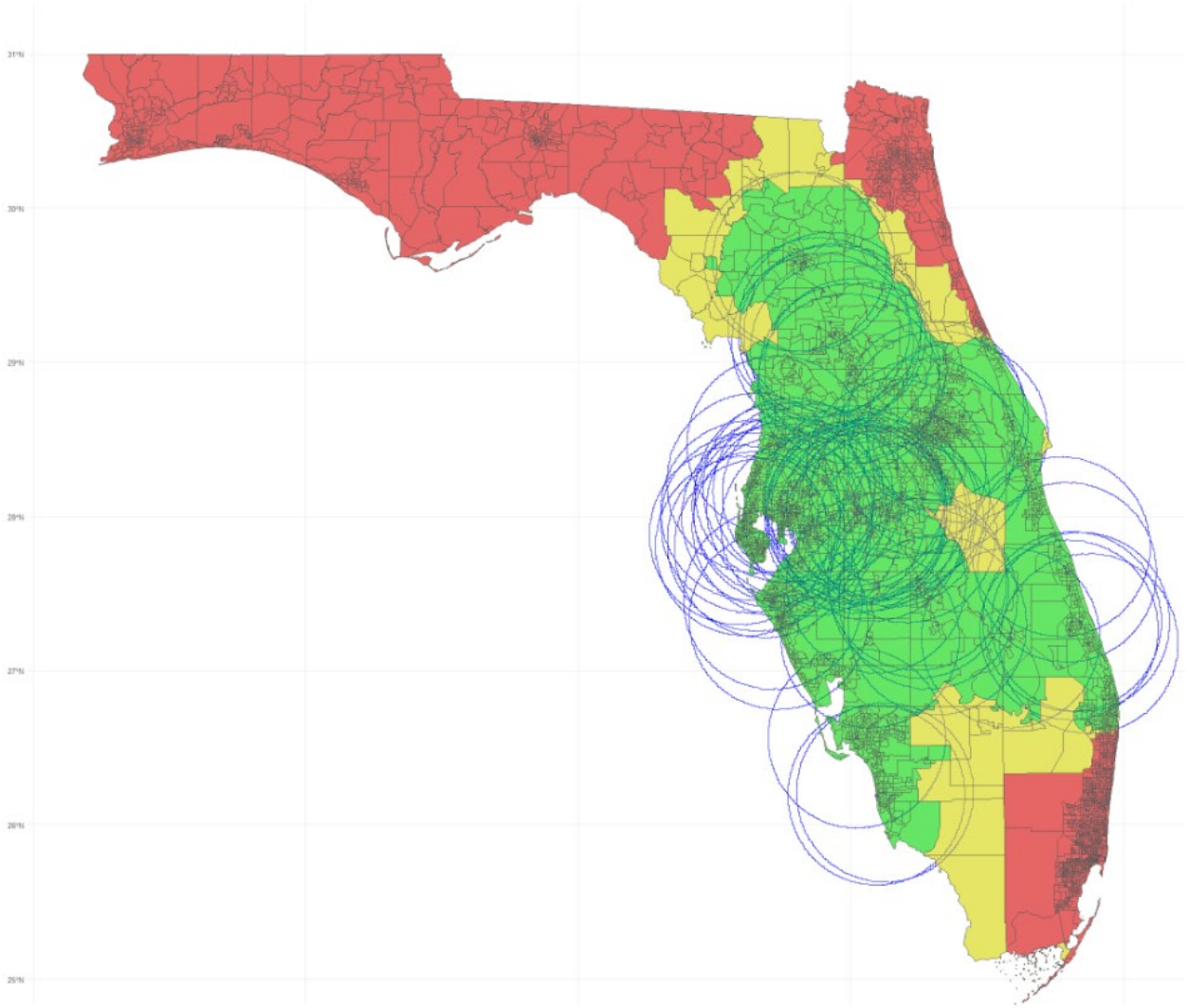

Figure 3: Map of Florida with 40 mile radius circles drawn around existing MidFloridas

Tracts that are not within 40 miles of an existing branch are shown in red, tracts that are partially within are shown in yellow, and fully within are shown in green.

4 Conclusion

To conclude, the twostep Heckman models that were created were able to predict the number of checking accounts for each census tract in the state of Florida. However, there is no way to tell if the Heckman model was able to fully account for the selection bias or not because there is no "truth" to test it against at this time. It can be difficult to prove that an instrumental variable is enough to fully counteract the selection bias and the instrumental variables used in the Heckman model may or may not have been enough. Still, the model proves that there is selection bias in their data and, therefore, their current method of creating predictions is not statistically sound. If in the future, MidFlorida wanted solely to rely on a Heckman model, they would have to be confident that whatever instrumental variables used are enough to counteract their selection bias.

5 Helpful Sources and Notation

"Best Practices in Empirical Research: Endogeneity - Selection Bias and the Heckman Two-Step Method." *YouTube*, YouTube, 15 Nov. 2021, www.youtube.com/watch?v=g9a_0gTMebA.

Online lecture given by Sarah Wolfolds of Cornell University.

CERTO, S. TREVIS, et al. vol. 37, Strategic Management Journal, 2016, *SAMPLE SELECTION BIAS AND HECKMAN MODELSIN STRATEGIC MANAGEMENT RESEARCH*, https://onlinelibrary.wiley.com/doi/epdf/10.1002/smj.2475?saml_referrer.

"The Heckman Sample Selection Model." *Rob Hicks*, 17 Sept. 2018, rlhick.people.wm.edu/stories/econ_407_notes_heckman.html.

Heckman Selection Model, www.stata.com/manuals/rheckman.pdf.

Toomet, Ott, and Arne Henningsen. "Sample Selection Models in R: Package sampleSelection." *CRAN*, cran.r-project.org/web/packages/sampleSelection/vignettes/selection.pdf.

Wolfolds, Sarah E., and Jordan Siegel. "Misaccounting for Endogeneity: The Peril of Relying on the Heckman Two-Step Method without a Valid Instrument." *Deep Blue Library Michigan University*, deepblue.lib.umich.edu/bitstream/handle/2027.42/147758/smj2995.pdf?sequence=1.

Appendix 5 – Market Saturation Model Design and Implementation

1. Data

The data used is US census data provided by MidFlorida. Each row describes a census tract in Florida, which are uniquely identified by the "geography" and "geographic_area_name" columns. Variables were generated to create population ratios for many of the variables. For example, the number of people with a bachelor's degree was divided by the total population for each row to create the bachelor's degree population ratio. Also, the natural log for each variable was taken. This allows for changes in variables to be viewed proportionally.

2. Multiple Linear Regression

To create the Poisson regression model described in the main report, a multiple linear regression model was first developed to determine which demographic variables were significant predictors. This linear model predicts the natural log of break-even number of financial institutions per capita, which is defined as:

$$Y = \frac{\text{Number of Financial Institutions} + 1}{\text{Total Population}} \qquad (1)$$

By estimating this instead of the number of financial institutions, the linear model can achieve more transferrable results in terms of which variables have the most predictive power for the Poisson model.

2.1. Model Selection

In Stata, the *v-select* command was used to perform variable selection on all available demographic variables. The predictors for the best three models (all of which the natural log was taken) are:

Model 1: Population density, square miles, bachelor's degree population ratio, elderly population ratio, renter occupied housing ratio

Model 2: Population density, square miles, bachelor's degree population ratio, elderly population ratio, renter occupied housing ratio, median household income, owner-occupied with a mortgage housing ratio

Model 3: Population density, square miles, bachelor's degree population ratio, elderly population ratio, renter occupied housing ratio, median household income, owner-occupied with a mortgage housing ratio, median home value, under 20 years old population ratio, worker population ratio, owner-occupied housing ratio

The model selection criteria include AIC, BIC, and the root mean square error (RMSE). AIC and BIC are both measures for prediction error, where BIC corrects for overfitting. Smaller values for each criterion are considered better. The summary of the results is in Fig 6.1 below.

Model	Number of Predictors	AIC	BIC	RMSE
1	5	8762.568	**8801.532**	0.59314
2	7	**8749.646**	8801.597	**0.59113**
3	11	8756.216	8842.637	0.59219

Fig 1: Model Selection Criteria Summary

From the values in Figure 1, Model 1 had the best BIC while Model 2 had the best AIC and RMSE and is very close to the best BIC. So, the predictors from Model 2 were selected to be used in the Poisson model.

3. Poisson Regression Model

Using the determined significant predictors, the final Poisson regression model was formulated in Equation 2 using robust standard errors.

$$\log(\widehat{Y}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \qquad (2)$$

Where

$$\begin{aligned} Y &= \text{Number of Financial Institutions} \\ X_1 &= \text{Log Population Density} \\ X_2 &= \text{Log Square Miles} \\ X_3 &= \text{Log Bachelor's Degree Population Ratio} \\ X_4 &= \text{Log Elderly Population Ratio} \\ X_5 &= \text{Log Renter Occupied Housing Ratio} \\ X_6 &= \text{Log Median Household Income} \\ X_7 &= \text{Log Owner Occupied with a Mortgage Housing Ratio} \end{aligned}$$

The coefficients and their standard errors are as follows.

$$\begin{aligned} \beta_0 &= -5.884,\ \text{SE} = 1.34 \\ \beta_1 &= 0.492,\ \text{SE} = 0.056 \\ \beta_2 &= 0.611,\ \text{SE} = 0.057 \\ \beta_3 &= 0.560,\ \text{SE} = 0.056 \\ \beta_4 &= 0.547,\ \text{SE} = 0.054 \\ \beta_5 &= 0.779,\ \text{SE} = 0.061 \\ \beta_6 &= 0.402,\ \text{SE} = 0.110 \\ \beta_7 &= -0.095,\ \text{SE} = 0.050 \end{aligned}$$

3.1. Example for Census Tract 110.18; Hillsborough County; Florida

The following calculations show a worked-out example for the census tract of one of the proposed sites. The demographic variables for Census Tract 110.18 are listed below.

$$
\begin{aligned}
Y &= \text{Number of Financial Institutions} = 1 \\
X_1 &= \text{Log Population Density} = \ln(5122.5) \\
X_2 &= \text{Log Square Miles} = \ln(1.031909) \\
X_3 &= \text{Log Bachelor's Degree Population Ratio} = \ln(0.1984487) \\
X_4 &= \text{Log Elderly Population Ratio} = \ln(0.1292092) \\
X_5 &= \text{Log Renter Occupied Housing Ratio} = \ln(0.8140233) \\
X_6 &= \text{Log Median Household Income} = \ln(53589) \\
X_7 &= \text{Log Owner Occupied with a Mortgage Housing Ratio} = \ln(0.1049869)
\end{aligned}
$$

Next, the values are plugged into the regression model to predict the equilibrium number of financial institutions.

$$
\begin{aligned}
\log(\widehat{Y}) = (-5.883905) + (0.4921156) \cdot \ln(5122.5) + (0.611329) \cdot \ln(1.031909) + \\
(0.5600428) \cdot \ln(0.1984487) + (0.546925) \cdot \ln(0.1292092) + (0.7793901) \cdot \ln(0.8140233) + \\
(0.4020964) \cdot \ln(53589) + (-0.0953013) \cdot \ln(0.1049869)
\end{aligned}
$$

$$
\log(\widehat{Y}) = 0.747
$$

$$
\widehat{Y} = e^{0.747} = 2.110 \text{ Financial Institutions}
$$

The prediction for the market equilibrium number of financial institutions is about 2, which is greater than what currently exists. This indicates that this census tract is possibly underserved.

4. Probabilities Using Poisson Distribution

From the Poisson regression model, an indicator for potential profitability was created using the Poisson distribution. The probability that the number of financial institutions at equilibrium is greater than what currently exists was calculated at each census tract. The probability mass function for a variable following a Poisson distribution is defined below (Davis et al., 2022).

$$
f_X(x) = (X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots \tag{3}
$$

Where

$$
\begin{aligned}
X &= \text{Number of Financial Institutions at Equilibrium} \\
\lambda &= \text{Expected (Fitted) Value}
\end{aligned}
$$

Since the value for lambda in a Poisson distribution is equal to the expected value, lambda is equal to the fitted values from the regression model.

Finally, the census tracts with the highest probabilities were given to the other sub-teams to do further investigation. The sub-teams would look into these specific census tracts, perform analysis using web scraping, traffic data, and spatial analysis to find potential locations.

4.1. Example for Census Tract 110.18; Hillsborough County; Florida

As in the previous section, the probability that the equilibrium number of financial institutions for Census Tract 110.18 is greater than what currently exists is worked out as an example.

The current number of financial institutions in this census tract is 1. So, to determine the probability that the equilibrium number is greater than 1, the probability must be calculated at 0 and 1.

$$P(X = 0) = \frac{2.110^0 \cdot e^{-2.110}}{0!} = 0.121$$
$$P(X = 1) = \frac{2.110^1 \cdot e^{-2.110}}{1!} = 0.256$$

These probabilities added is equal to the probability that the equilibrium number of financial institutions is less than or equal to 1. So, the sum of these two probabilities is subtracted from 1.

$$\begin{aligned} P(X > 1) &= 1 - (P(X = 0) + P(X = 1)) \\ &= 1 - (0.121 + 0.256) \\ &= 0.623 \end{aligned}$$

Given the expected value of 2.11 from the regression model, the probability that the equilibrium number of financial institutions for Census Tract 110.18 is greater than 1 is 0.623. This is a fairly high probability; however, this census tract was only chosen after being confirmed by the analysis from the other sub-teams.

5. References

Davis, M., Hitt, B., Horton, H., Pruitt, K., Warner, B. (2022). *Computational Probability and Statistics*. Github.https://github.com/DS-USAFA/Computational-Probability-and-Statistics/blob/main/12-Named-Discrete-Distributions.Rmd

Appendix 6 – Web Scraping Report

1 Introduction

MIDFLORIDA Credit Union requested new techniques towards solving their problem of selecting a location for a new branch site. The alternative methods sub team used Web Scraping as a new tool in data collection for MIDFLORIDA. The purpose is to use Web Scraping and a manual approach to determine a convenient location for customers based on the businesses they frequent. Time equals money; therefore, most people will choose to be a member of a bank in the most convenient location for them. A convenient location could be by their favorite restaurant (maybe to get cash for a tip), next to the pharmacy where they get their prescriptions, or next to the grocery store where they do all their shopping. Most people will not make a special trip just to go to the bank, so to make a branch successful, it must be by common businesses. The businesses that were selected based on the data collection were then added to the visualization map alongside the other sub team data including demographics, market saturation, and traffic analysis to help determine a new branch site location.

2 What is Web Scraping?

Web scraping is a technique used to collect content and data from websites using bots that are programmed to visit websites, grab the relevant pages, and extract useful information. Web scraping works by making an HTTP request to a server, extracting and parsing the website's code, and saving the relevant data locally. There are many different categories of Web scrapers, but for this project specifically scraping extensions were used. Scraping extensions allow the user to scrape web pages as they navigate in real-time and are the simplest web scraping tools. Typically, free and requires little prior knowledge of web scraping.

3 Selecting Bank Locations

To begin this project, the team first needed to decide on which banks to consider as part of the dataset. To do this, the team used the model file where a column of ones and zeros was added to indicate whether the MIDFLORIDA branch had an average contribution that was greater than or equal to one-hundred thousand. Due to the time and resources limit, the team picked 15 successful MIDFLORIDA branches (greater than 100K) and 15 less successful ones (less than 100K). The team also randomly selected five Suncoast Credit Union locations, five Truist locations, and five Wells Fargo locations to give a total of 45 bank locations that they would be comparing. They did this by using Google Maps and saving the branches to a list to make sure they had a wide variety over the state of Florida as seen in figure 1.

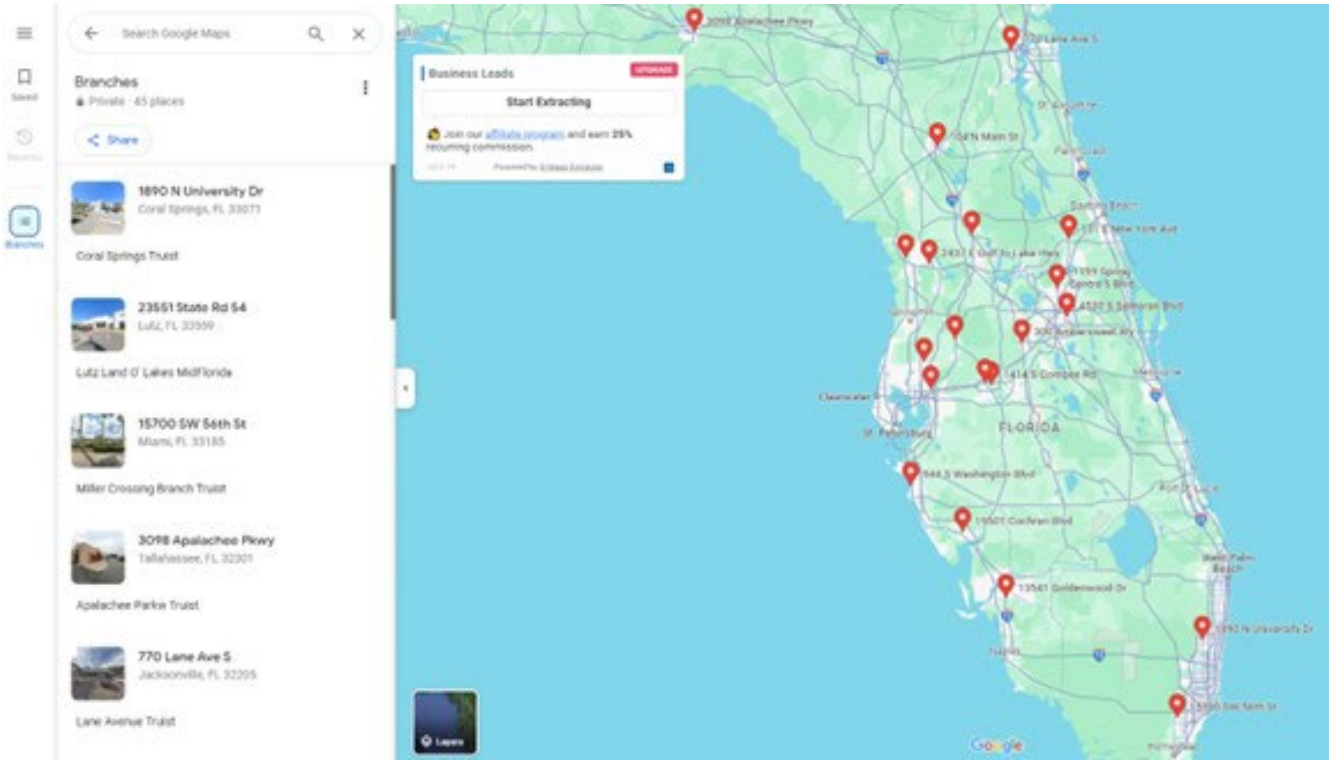

Figure 1: Adding all branches to Google Maps for visualization.

4 Solution Paths

In the ideal scenario, there would be a website that exists where someone could enter an address and a radius, and the website would output a list of all franchised businesses within that radius. Unfortunately, during the team's first semester of research they were unable to find such an application. Instead, they found many applications that work based off categories since Google Map uses categories to search for businesses. The team decided to use one of these applications as the automated approach, and another website where there is an option to select the radius but not the category as the manual approach. The reason the team decided to use both an automated and manual approach, is because the manual approach is slower but more precise (since the radius is specified), while the automated approach is faster and gathers less data (limited to categories).

5 Manual Approach

The manual approach involved using a website called Google Map Developers using the "Draw a Circle Tool". The first step was entering the bank location address and then deciding to set the radius to two miles. The second step is clicking the "Zoom to Address" button along with the "New Circle" button. The demonstration of the first two steps can be seen in figure 2. As a note when using this website, in a few cases sometimes the center of the circle was not at the correct address, so make sure to zoom in all the way to verify when using this tool. The final step was zooming into the top of the circle and working left to right/top to bottom recording the franchises in an Excel document. The team repeated these steps for all 45 bank locations. The data was broken up into three parts (15 locations each): less successful MIDFLORIDA, successful MIDFLORIDA, and other banks (Truist, Suncoast, Wells Fargo).

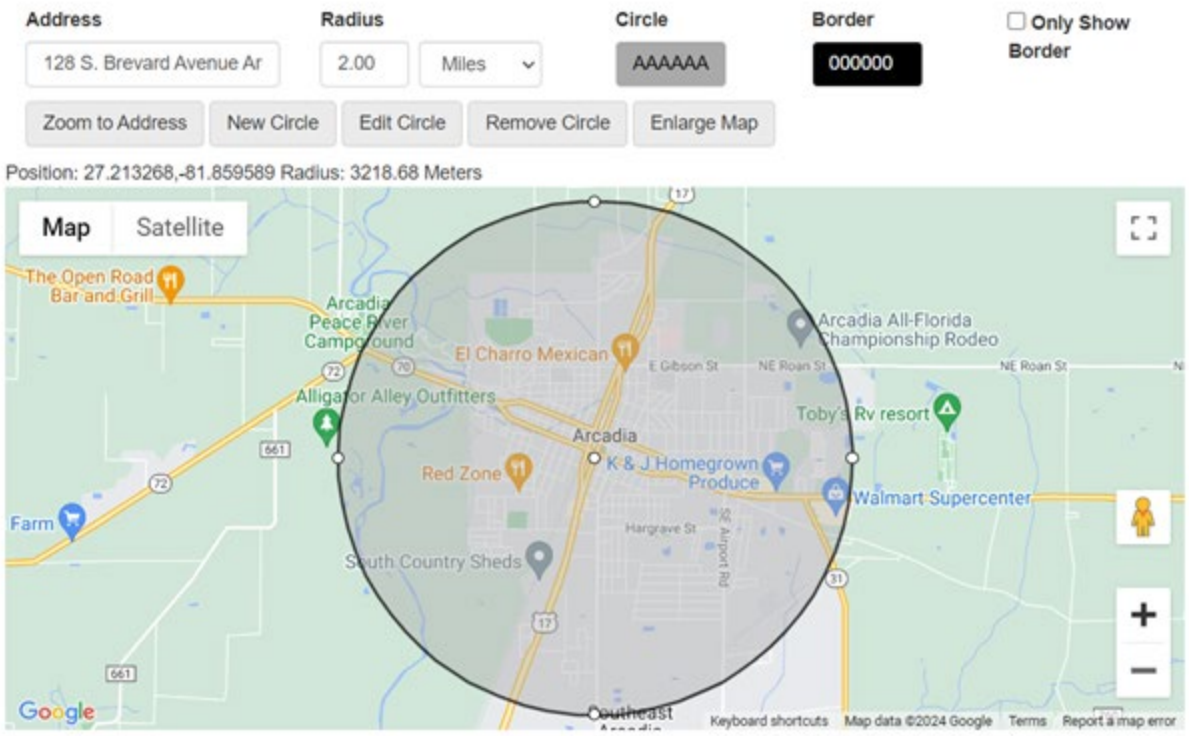

Figure 2: Demonstrating the manual method using Google Map Developers.

6 Automated Approach

The automated approach involved using a Web scraping extension called G Maps Extractor and adding it to the Chrome browser. This extension works when a category is searched in Google Maps, the extension can extract the first ten search results to an Excel file. Unfortunately, the free version only allows 1,000 results for free each month. Therefore, the team decided to search two categories of "Stores" and "Restaurants" for all 45 bank locations leading them to scrape 900 results to leave room for error. Once the team downloaded the extension, the first step was opening Google Maps and typing in the address of a branch. The second step was clicking the "Nearby" button on Google Maps to try and get closer businesses. The third step was entering the category "Restaurants" and clicking the "Start Extracting" button on the extension. The demonstration of the third step can be seen in figure 3. The fourth step was clicking the "Export Detailed List" button and then the "Reset" button to reset the extension. The final step was entering the category "Stores" and clicking the "Start Extracting", "Export Detailed List", and "Reset" buttons. The team then repeated these steps for all 45 bank locations. The data was broken up into three parts (15 locations each): less successful MIDFLORIDA, successful MIDFLORIDA, and other banks (Truist, Suncoast, Wells Fargo).

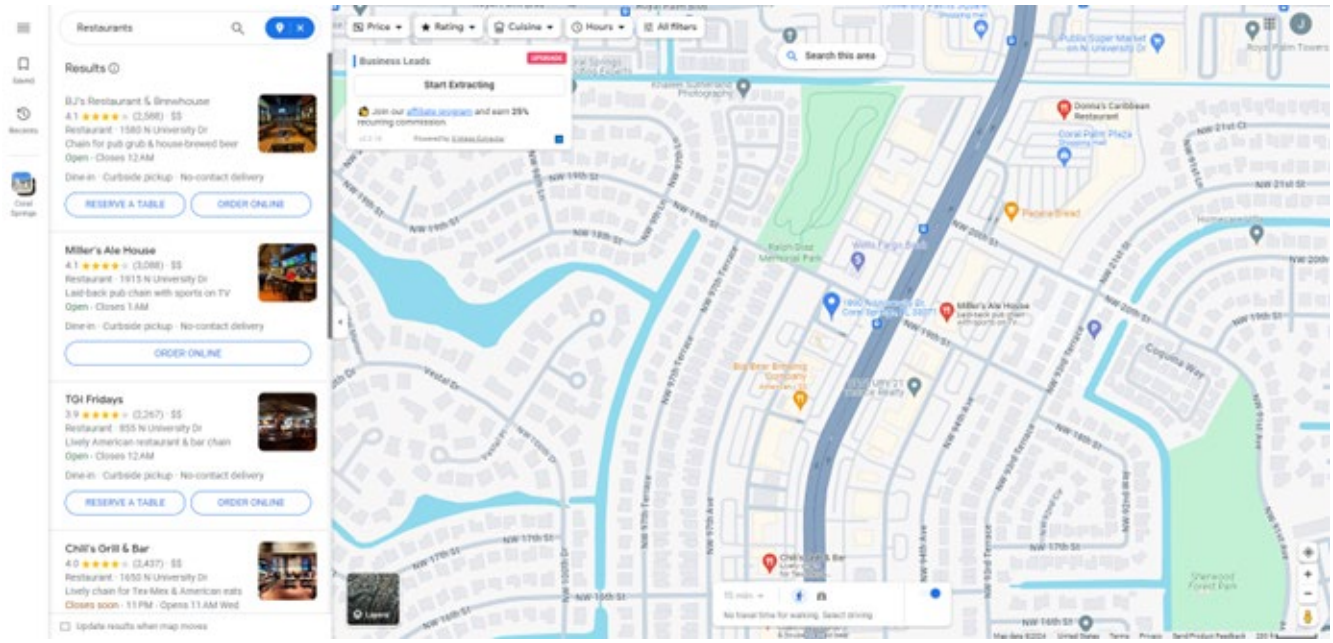

Figure 3: Demonstrating the automated method using G Maps Extractor browser extension.

7 Analyzing Results

Once all the results were analyzed individually for the automated and manual methods it was time to pull it all together. The team first created a file containing the automated and manual results for the successful and less successful MIDFLORIDA branches and the other banks. These results include all the businesses that were by more than one branch and includes the business name, category, and number of branches that business is by out of 15, as well as the percentage. Once the team created this file, they found it difficult to decide on what businesses to select to add to the final visualization map. After discussing with Dr. Dewey, the team determined that the most important businesses were the businesses that were by all banks (successful MIDFLORIDA, Wells Fargo, Suncoast, Truist). In addition, the businesses that were found to be by successful MIDFLORIDA branches, but not by less successful MIDFLORIDA branches. The team then created a file summarizing based on the automated or manual methods and listing the top three results. By looking at businesses that are by all banks, it can help narrow down the census tracts the team is looking at, by not focusing on census tracts that do not have these staple businesses. By looking at businesses that are by successful MIDFLORIDA but not by less successful MIDFLORIDA, the team can help narrow down a specific area/intersection that contains the businesses that could make MIDFLORIDA more likely to be successful.

8 Selected Businesses

The team used the summarized results file to decide which businesses to add to the visualization map. The first business they did was Publix since Publix had been brought up many times by the sponsor as a business they consider when looking at a location. Based on the data the team collected, they then had the evidence to support this as Publix was by 87% of all banks. Luckily, the sponsor was able to get all the locations of Publix in the US from a collegue, so the team just had to create a file with only the Florida locations. The next business the team tried to do was McDonald's since it was by 100% of all banks but unfortunately, they were not able to find all the locations in Florida available online. The next business they did was Dollar General using their website and another Web scraping browser extension Webscraper.io. Webscraper.io is by far the simplest and easiest Web scraper browser extension the team has found. The Chrome plugin is completely free and there are

many YouTube videos along with documentation on how to use the extension. The next business they did was Cheddar's Scratch Kitchen since it was by 53% of successful MIDFLORIDA branches and by no less successful. The next business they tried to do was CVS which the team was able to find the website that listed all the locations, but unfortunately could not get the Web scraper browser extension to work on this site due to the backend of how the website was made. The next business they did was Advance Auto Parts since it was by 80% of all banks. The final business they did was Beef 'O' Brady's since it was by 47% of successful MIDFLORIDA branches and by no less successful. To use Webscraper.io on Google Chrome select the top right corner drop down and select More tools > Developer tools and there will be a tab for Web Scraper. For the Dollar General, Cheddar's, Advance Auto, and Beef 'O' Brady's websites, to get the address data that the team wanted, first they had to click the link to that specific location's website as it wasn't listed on the main page. Luckily, Webscraper.io has a function for this and they were able to find a YouTube video to demonstrate this process of using the Pagination Selector. While as they stated with CVS this extension is not able to work for every website, the team highly recommends this free tool. For further explanation along with step-by-step tutorials on how to get started, visit their website for documentation. Once the data was exported from these sites, the addresses needed to be converted to latitude/longitude to add the businesses to the visualization map. Thankfully, the team's sponsor was able to do this as the team was unable to find a free way to do the conversion. Because of the way some of the websites were set up, the extraction of data did not come out into the specific rows and columns that the team wanted. To fix this they used an Excel add-in called Ablebits where they specifically used the "Merge Duplicates" and "Split Names" features.

9 Further Recommendations

As mentioned in the Solution Paths section of this report, the ideal application for this problem does not exist yet. One recommendation is to find a way to create this application or continue looking into other Web scraping software as they advance. One website the team did find promising was Outscraper which allows you to enter an address and search all possible categories and extract all the results. However, this website does not have the option to specify a radius and costs money. Another recommendation is to compare the businesses around the rest of the successful and less successful MIDFLORIDA locations along with more locations for other banks. The final recommendation is to use a tool like G Maps Extractor or Outscraper where once a potential location is determined, the sponsor can further investigate the businesses around that location using these tools.

10 Conclusion

Overall, Web scraping is a powerful tool that can be used in many ways and can be used for many applications. Comparing the businesses around potential sites is a major component in the potential profitability of that site. Possibly learning more about MIDFLORIDA customers and when they tend to stop by the bank could also be useful in solving this problem.

11 References

Chouinard, Jean-Christophe. "Web Scraping Tutorial (with Examples)." *JC Chouinard*, 7 May 2023, www.jcchouinard.com/web-scraping/.

Hillier, Will. “What Is Web Scraping? A Complete Beginner’s Guide.” *CareerFoundry*, 13 Aug. 2021, https://careerfoundry.com/en/blog/data-analytics/web-scraping-guide/.

Appendix 7 – Heat Maps and Site Recommendations

1. Heat Maps

Heat maps were used to combine the different components of determining branch site location for MidFlorida. Doing so provided an interactive, visual display that MidFlorida can use in their analysis. Figure 1 demonstrates an example of the interactive heat maps created for analysis. To find a viable site, the probability of an area being underserved is examined in conjunction with the predicted number of checking accounts. Promising census tracts are further examined using heat maps that show features that are instrumental in determining if a site will be profitable.

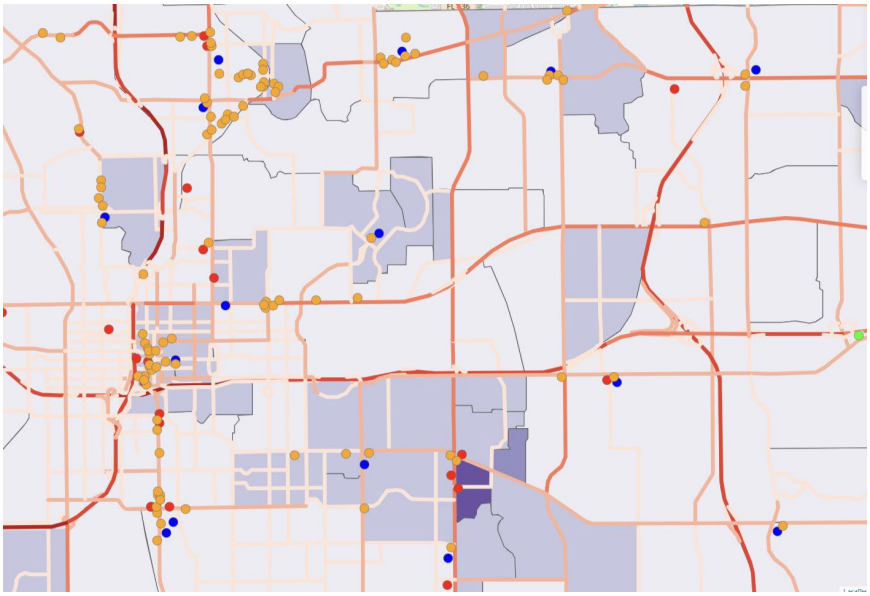

Figure 1: Area of selected location, showing population density, Publix locations, and competitor locations. Blue dots are Publix locations, orange dots are banks, red dots are other credit unions, and red lines depict the annual average daily traffic (AADT).

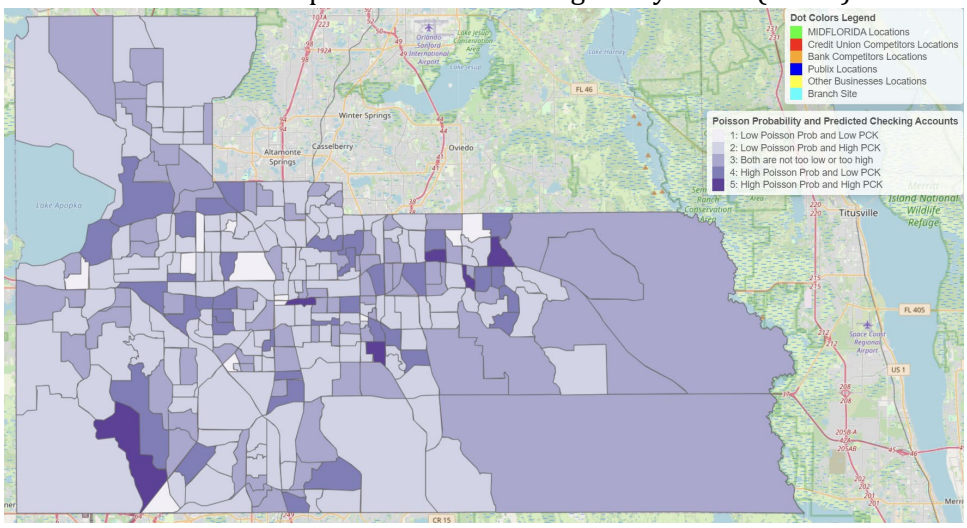

Figure 2: Poisson Probability and Predicted Checking Account heat map. The first heat map to be used as it provides the team the information about which census tracts have the highest probability of being underserved and the highest predicted checking accounts.

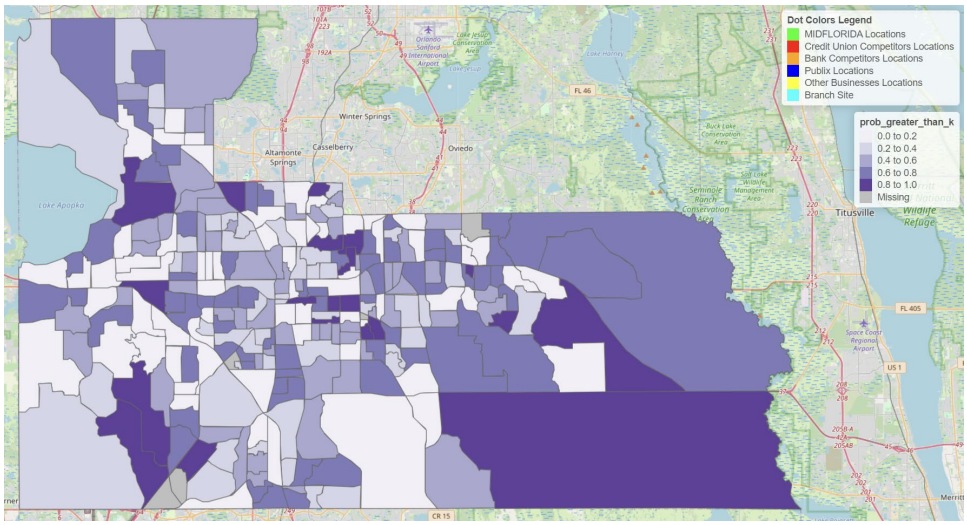

Figure 3: Poisson Probability heat map. Darker shade represents a higher probability of being underserved.

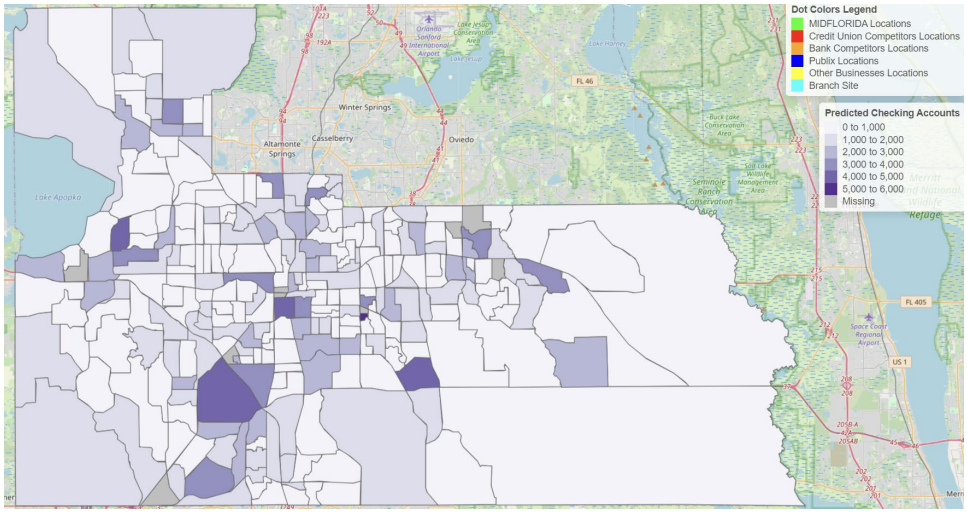

Figure 4: Predicted Checking Account heat map. Darker shade represents a higher predicted checking account.

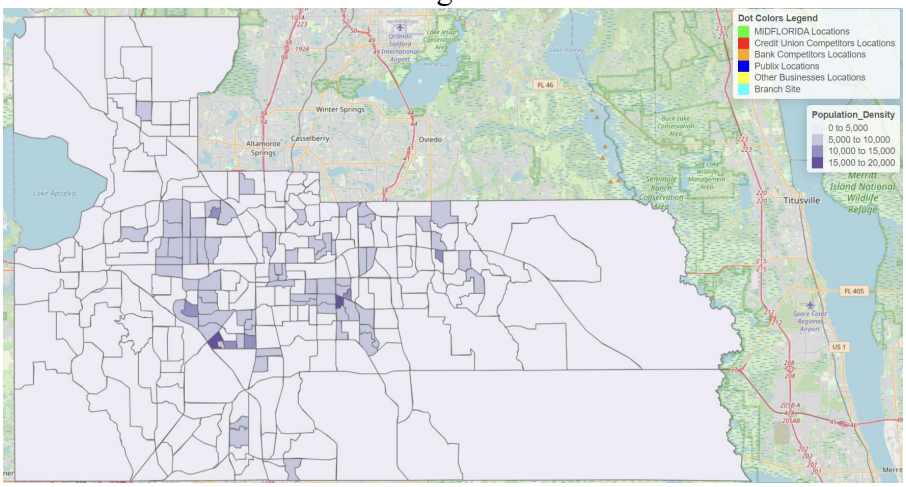

Figure 5: Population Density heat map. Darker shade represents a higher population density.

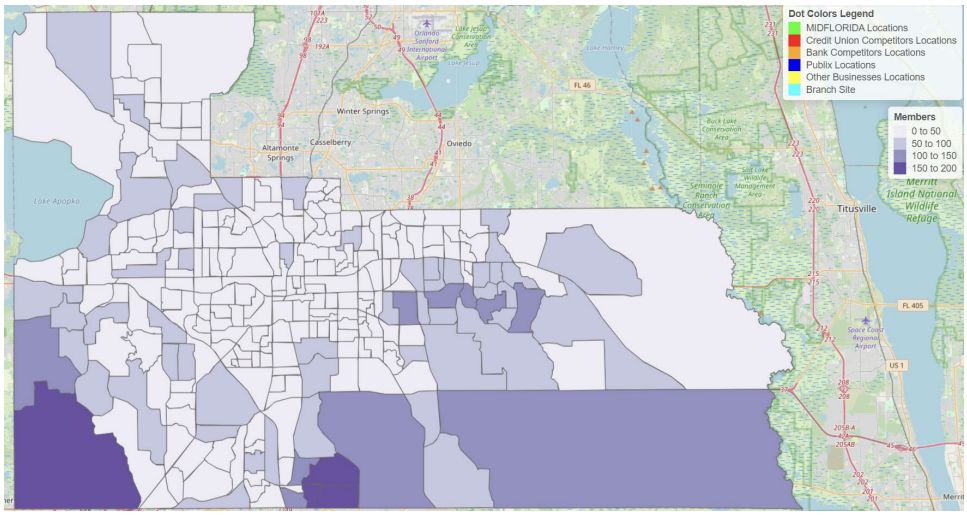

Figure 6: MidFlorida Members heat map. Darker shade represents a higher MidFlorida members population.

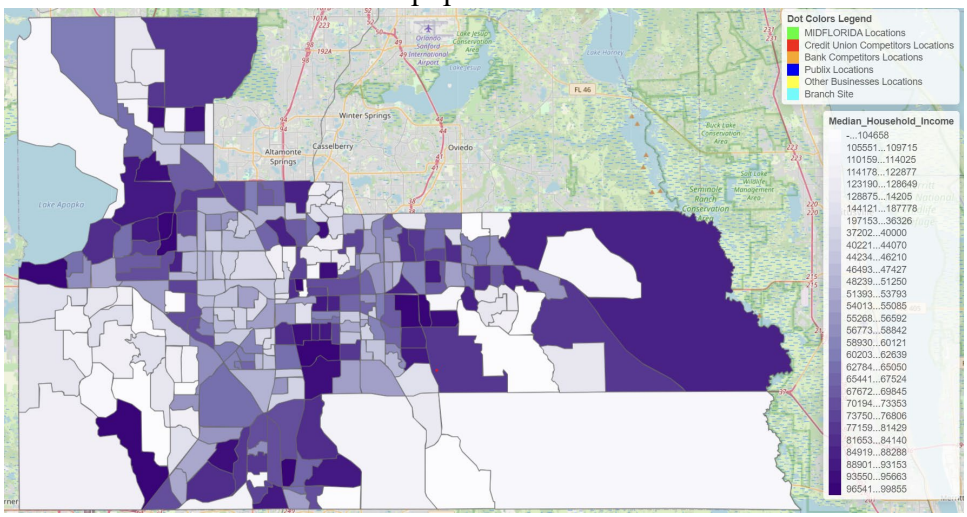

Figure 7: Median Household Income heat map. Darker shade represents a higher median household income.

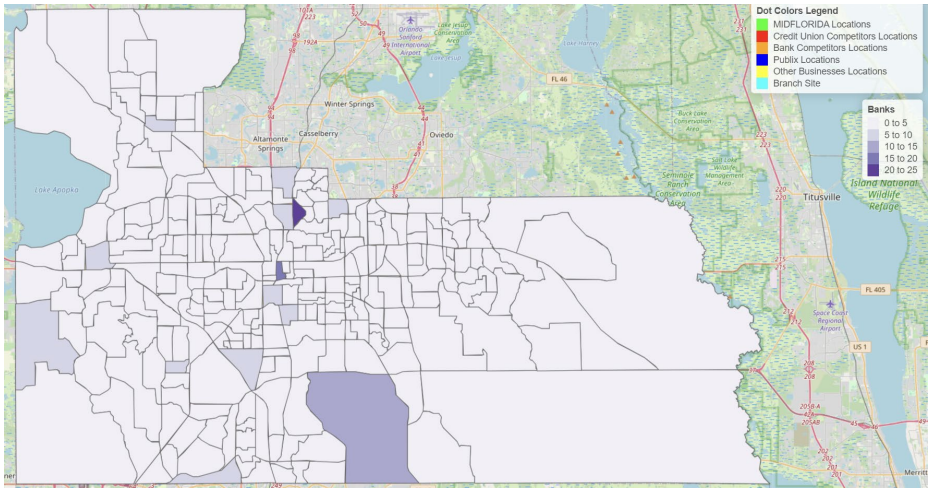

Figure 8: Banks heat map. Darker shade represents a higher count of banks.

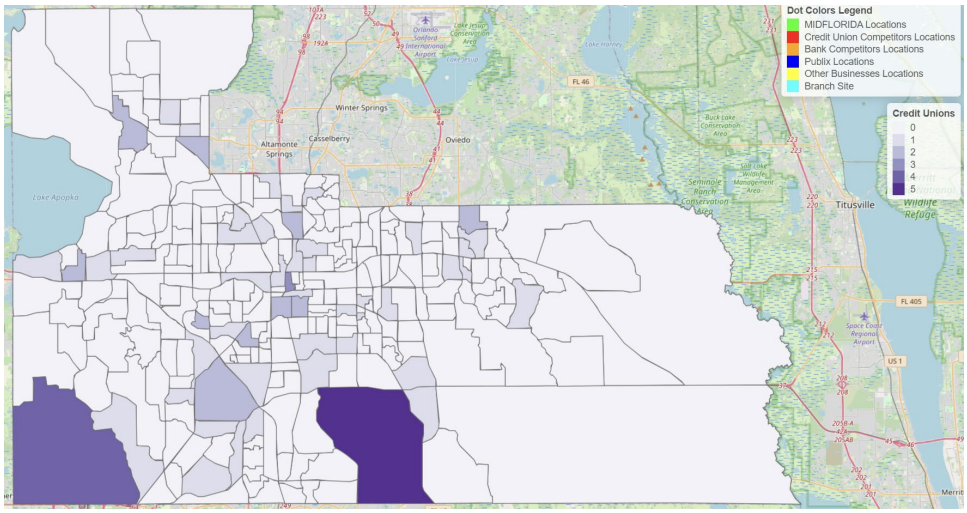

Figure 9: Credit unions heat map. Darker shade represents a higher count of credit unions.

2. Site Suggestions

The first site suggestion is at 1477 McMullen Booth Road. This is in census tract 268.16 in Pinellas County. The probability of this area being underserved is .65. The predicted checking accounts is 779. The average annual daily traffic is 40,500. The population density is 3,148. The median household income is $98,528. The location is at a former Chase branch location, and there are no competitors nearby.

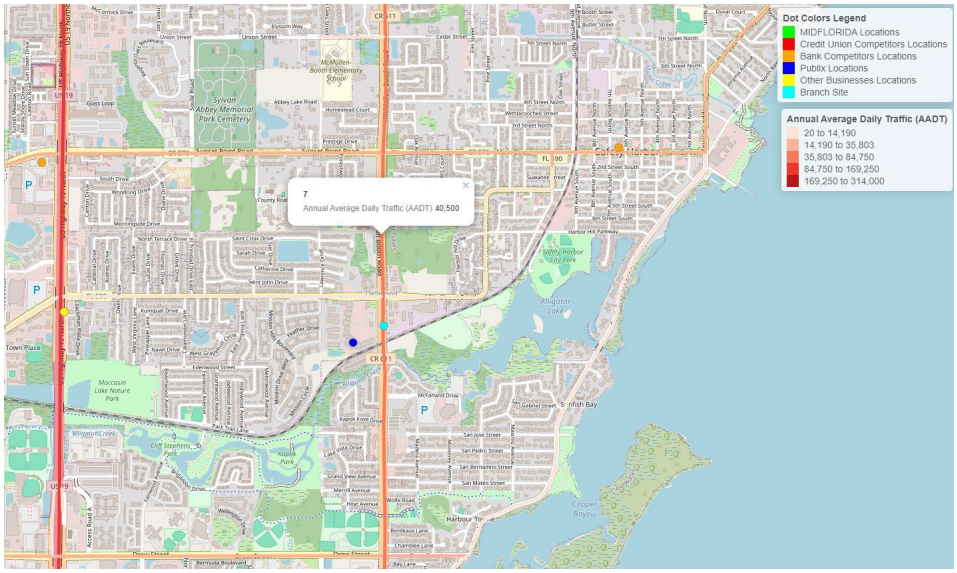

Figure 10: First site recommendation

The second location is at 15302 Amberly Dr, in census tract 110.18, in Hillsborough County. The probability of this area being underserved is .62. The predicted checking accounts is 799. The average annual daily traffic is 51,000. The population density is 5,122. The median household income is $53,589. The location is at a former Bank of America branch location, and there are a Publix and Truist Bank on the same road.

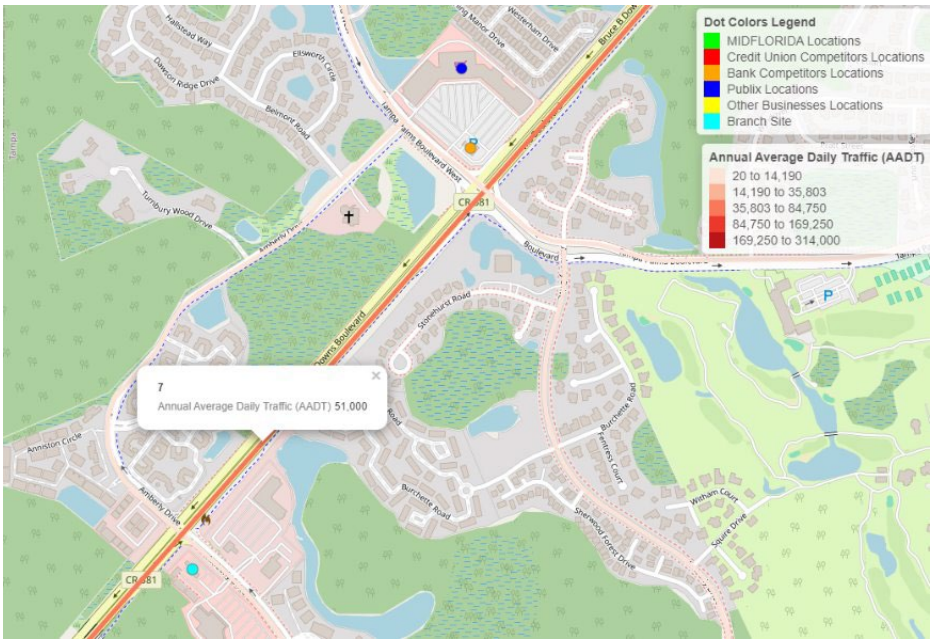

Figure 11: Second site recommendation

The third location is at 1870 Aloma Ave, in census tract 160.02, in Orange County. The probability of this area being underserved is .41. The predicted checking accounts is 976. The average annual daily traffic is 34,500. The population density is 3,138. The median household income is $228,848. The location is at a former Chase Bank branch location, and there is a Publix and other banking institutions nearby (but no credit unions). It is also located on the corner of an intersection.

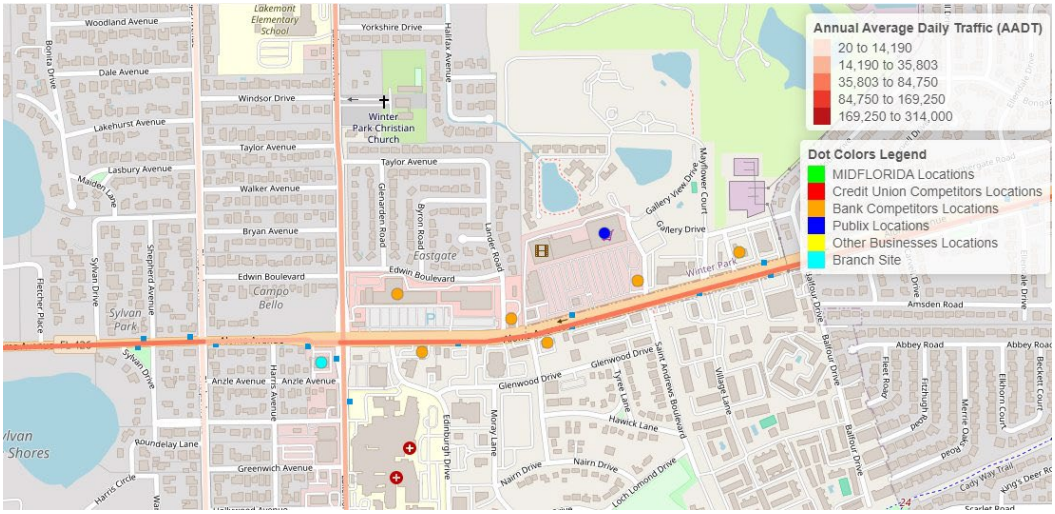

Figure 12: Third site recommendation